Information Retrieval untuk Identifikasi Pasal Hukum yang Mendukung Permasalahan Keuangan di Indonesia

Cuthbert Young, Gizha Pradipta, Shidqy Baihaqy El Muhammady, Tika Dian Pangastuti, Venedict Grinaldy Prasetyo

Abstract

Background: Pertumbuhan eksponensial dokumen regulasi keuangan di Indonesia menghadirkan tantangan signifikan dalam sistem temu kembali informasi (*Information Retrieval*), khususnya bagi pengguna non-ahli. Metode pencarian konvensional sering kali gagal menjembatani kesenjangan kosa kata (*vocabulary mismatch*) antara kueri pengguna yang bersifat informal dan deskriptif dengan terminologi hukum (*legalese*) yang baku dan kaku dalam dokumen peraturan.

Objective: Penelitian ini bertujuan untuk mengembangkan dan mengevaluasi kinerja sistem *Statutory Retrieval* yang mampu mengidentifikasi pasal-pasal relevan pada korpus peraturan keuangan menggunakan arsitektur *Two-Stage Retrieval*. Fokus utama penelitian adalah membandingkan efektivitas pendekatan leksikal, semantik, hibrida, dan *re-ranking* dalam menangani variasi kueri hukum yang kompleks.

Methods: Kami menerapkan strategi eksperimen komparatif menggunakan lima pendekatan model: (1) TF-IDF (n-gram), (2) BM25 (Leksikal), (3) Dense Retrieval menggunakan model *pre-trained* intfloat/multilingual-e5-base yang dilatih dengan *contrastive learning*, (4) Hybrid Retrieval yang menggabungkan skor BM25 dan Dense menggunakan algoritma *Reciprocal Rank Fusion* (RRF), dan (5) Re-ranking menggunakan *Cross-Encoder* BAAI/bge-reranker-base. Berbeda dengan penelitian sebelumnya yang menggunakan *ground truth* tunggal, penelitian ini menerapkan strategi evaluasi *Expanded Ground Truth* dengan mekanisme penilaian berbasis Overlap Coefficient. Metode ini memungkinkan sistem untuk mendeteksi relevansi substansi pasal meskipun terdapat variasi format atau duplikasi dokumen yang umum terjadi pada data hukum.

Results: Hasil evaluasi pada metrik nDCG@K dan Recall@K menunjukkan bahwa pendekatan *Dense Retrieval* (E5) secara signifikan mengungguli metode leksikal dalam menangkap makna semantik kueri konseptual. Lebih jauh, integrasi tahap kedua menggunakan BGE-Reranker terbukti memberikan performa terbaik dalam kualitas peringkat (*ranking quality*), secara efektif menempatkan pasal yang paling relevan di posisi teratas. Sementara itu, pendekatan Hybrid (RRF) terbukti sebagai metode yang paling tangguh (*robust*) dalam menjaga stabilitas *Recall*, menyeimbangkan presisi kata kunci eksak untuk nomor pasal dan pemahaman konteks untuk deskripsi kasus.

Conclusion:.Penelitian ini menyimpulkan bahwa arsitektur *Retrieve-and-Rerank* yang menggabungkan kekuatan pencarian hibrida dan penalaran mendalam *Cross-Encoder* adalah solusi paling optimal untuk sistem pencarian hukum di Indonesia. Penggunaan metrik evaluasi berbasis *overlap* juga direkomendasikan untuk meningkatkan keadilan penilaian (*fairness*) pada korpus dokumen yang memiliki tingkat redundansi tinggi.

Keywords: Information Retrieval, Statutory Retrieval, Multilingual-E5, BGE-Reranker, Hybrid Search, Reciprocal Rank Fusion, Hukum Keuangan.

I. Introduction

Jumlah peraturan dan dokumen hukum yang diterbitkan pemerintah Indonesia terus meningkat setiap tahun, mulai dari undang-undang, peraturan pemerintah, hingga regulasi teknis yang diunggah ke berbagai portal hukum nasional. Pertumbuhan dokumen yang begitu cepat membuat proses menemukan informasi hukum yang tepat menjadi semakin sulit dan memakan waktu. Dalam situasi ini, teknologi *Natural Language Processing* (NLP) menjadi sangat penting karena mampu membantu komputer memahami dan memproses bahasa manusia secara otomatis (Shoukat Ali & Shandilya, 2021). Selain itu, teknik *text mining* mampu mengekstraksi pola dan pengetahuan dari kumpulan teks berskala besar, sehingga sangat relevan digunakan untuk menangani dokumen hukum yang kompleks dan tidak terstruktur (Kaur & Kaur, 2019).

Pada konteks peraturan keuangan di Indonesia, tantangan ini semakin terlihat karena dokumen hukum keuangan cenderung tebal, bersifat teknis, dan tersebar dalam berbagai regulasi berbeda. Proses pencarian manual atau pencarian sederhana berbasis kata kunci kerap tidak

mampu menemukan pasal yang benar-benar relevan dengan situasi tertentu. Taylor (2019) menegaskan bahwa sistem informasi hukum tradisional sering kali gagal memahami konteks dan struktur bahasa hukum, sehingga hasil pencarian menjadi kurang akurat. Kondisi ini tidak hanya menyulitkan analis hukum atau peneliti, tetapi juga masyarakat umum yang mengalami kasus keuangan dan perlu mencari pasal yang sesuai untuk memahami hak, kewajiban, atau landasan hukum atas permasalahan yang mereka hadapi.

Pentingnya penelitian ini terlihat dari kebutuhan untuk menemukan dasar hukum yang tepat dalam penyelesaian berbagai permasalahan keuangan baik terkait pengelolaan anggaran negara, layanan lembaga keuangan, maupun persoalan individu seperti sengketa transaksi, penipuan digital, atau kredit bermasalah. Sistem *information retrieval* modern mampu menyediakan mekanisme pencarian pasal hukum yang lebih cepat dan akurat melalui pemanfaatan algoritma NLP dan metode pencarian terstruktur (Hong et al., n.d.). Pendekatan ini tidak hanya meningkatkan efisiensi analisis hukum oleh para profesional, tetapi juga membantu individu yang mengalami masalah keuangan untuk mengetahui pasal yang relevan dengan kasus mereka. Dengan demikian, pengembangan IR dalam bidang hukum keuangan menjadi langkah strategis untuk meningkatkan aksesibilitas dan kualitas informasi hukum di Indonesia.

Meskipun potensi penerapannya besar, pengembangan sistem pencarian hukum menghadapi tantangan teknis yang signifikan dari sisi karakteristik data dan kesenjangan semantik. Dokumen hukum umumnya didistribusikan dalam format PDF dengan tata letak visual yang kompleks, seperti keberadaan *header*, *footer*, dan struktur tabel yang rumit sehingga sering kali gagal diekstrak secara utuh oleh alat *parsing* standar (Adhikari & Agarwal, 2025). Oleh karena itu, diperlukan teknik ekstraksi elemen spesifik yang mampu mentransformasi dokumen regulasi yang tidak terstruktur menjadi format basis data yang bersih dan siap diproses (Jiang & Li, 2023).

Selain tantangan format, kesenjangan kosa kata (*vocabulary mismatch*) menjadi hambatan utama ketika istilah keluhan keuangan yang digunakan masyarakat awam berbeda drastis dengan terminologi baku dalam undang-undang. Untuk mengatasi hal ini, pendekatan *Automatic Query Expansion* (AQE) dapat diterapkan guna memperkaya kueri pengguna dengan istilah sinonim yang relevan secara otomatis (Kulkarni & Kale, 2021), sementara pemanfaatan *Latent Semantic Analysis* (LSA) dapat membantu sistem menangkap makna implisit di balik teks yang ambigu (Li, 2024; Rezvani et al., 2023). Lebih jauh, penelitian terbaru menunjukkan bahwa akurasi pencarian dapat ditingkatkan secara signifikan melalui metode hibrida yang menggabungkan kekuatan statistik TF-IDF, dengan mempertimbangkan faktor kebaruan istilah atau *term-recency*, dan pemahaman konteks mendalam dari model berbasis *Deep Learning* seperti BERT (Aprilio et al., 2025; Marwah & Beel, 2021). Integrasi berbagai pendekatan inilah yang diperlukan untuk membangun sistem identifikasi pasal hukum yang tangguh dan relevan.

II. Related Works

Information retrieval pada dokumen hukum memiliki karakteristik yang berbeda dibandingkan korpus umum. Dokumen hukum memiliki teks panjang, kaya terminologi, dan sering punya struktur formal sehingga kebutuhan *retrieval* bukan sekadar “mirip kata”, tetapi juga relevan secara makna dalam konteks norma. Literatur merangkum bahwa legal IR menghadapi tantangan skalabilitas, kompleksitas semantik, serta kosakata domain-spesifik yang dapat menurunkan performa metode yang hanya mengandalkan pencocokan kata (Souza et al., 2021). Selain itu, terminologi hukum yang khusus dan fenomena polisemi menuntut strategi yang lebih adaptif agar pencarian tetap efektif.

Penelitian Information Retrieval (IR) pada domain hukum mengalami perubahan penting dalam beberapa tahun terakhir. Metode pencarian berbasis kata kunci seperti BM25 mulai ditinggalkan karena tidak mampu mengatasi vocabulary mismatch, yaitu kondisi ketika istilah dalam kueri berbeda dari terminologi baku dalam dokumen hukum (Thakur et al., 2021). BM25 banyak dilaporkan sebagai baseline yang robust dengan *recall* tinggi pada retrieval legislatif (Verma et al., 2023). Selain BM25, terdapat model leksikal lain seperti TF-IDF. TF-IDF efektif untuk pencarian berbasis kata kunci, tetapi memiliki keterbatasan mendasar ketika *query* dan dokumen tidak berbagi token yang sama akibat sinonimi/polisemi

Untuk menjawab tantangan ini, pendekatan dense retrieval menggunakan arsitektur Bi-Encoder berbasis Transformer menjadi standar baru. Studi pada benchmark BEIR menunjukkan bahwa model dense retrieval yang dilatih dengan contrastive learning dapat memahami makna semantik dengan jauh lebih baik daripada metode leksikal, terutama dalam skenario zero-shot pada domain khusus seperti hukum (Thakur et al., 2021). Selain itu, model multibahasa juga terbukti efektif untuk menangani variasi bahasa, dengan keunggulan bahwa vektor dokumen dapat dihitung sebelumnya untuk mempercepat proses pencarian skala besar (Bonifacio et al., 2021).

Meskipun Bi-Encoder menawarkan kecepatan tinggi, representasi dokumen dalam bentuk satu vektor sering kali belum cukup untuk menangkap interaksi kompleks antara kueri dan dokumen. Oleh karena itu, banyak sistem modern mengadopsi arsitektur multi-stage retrieval atau Retrieve-and-Rerank. Pada tahap pertama, Bi-Encoder digunakan untuk mendapatkan kandidat dokumen. Selanjutnya, model Cross-Encoder digunakan sebagai reranker yang memproses kueri dan dokumen secara bersamaan sehingga dapat memberikan skor relevansi yang lebih akurat (Pradeep et al., 2021). Penelitian menunjukkan bahwa strategi ini, terutama jika menggunakan model yang dilatih pada dataset besar seperti MS MARCO, mampu meningkatkan metrik nDCG secara signifikan dalam tugas penemuan pasal hukum (Louis et al., 2022). Dengan cara ini, sistem dapat mempertahankan recall tinggi dari tahap awal sekaligus meningkatkan kualitas peringkat akhir.

Selain aspek pencarian, tantangan lain dalam sistem informasi hukum adalah mengekstrak jawaban yang tepat dari dokumen panjang dan berstruktur kompleks. Hal ini mendorong integrasi teknik Question Answering (QA) ke dalam pipeline IR, membentuk arsitektur Retriever-Reader. Survei terbaru mengenai penggunaan Large Language Models (LLM) di bidang hukum menunjukkan bahwa model berbasis Transformer, seperti XLM-RoBERTa dan Legal-BERT, mampu melakukan ekstraksi informasi dan penalaran hukum secara akurat jika dokumen dibagi ke dalam segmen yang sesuai (Curran et al., 2023). Lebih jauh, benchmark LexGLUE memperlihatkan bahwa model yang melalui domain-adaptive pretraining pada korpus hukum menunjukkan kinerja yang lebih baik dalam memahami bahasa hukum (legalese) dibandingkan model umum (Chalkidis et al., 2021). Dengan demikian, penggabungan teknik retrieval semantik dan kemampuan machine reading comprehension menjadi fokus utama dalam pengembangan sistem pencarian regulasi yang lebih komprehensif.

Meskipun model berbasis *Cross-Encoder* unggul dalam akurasi, biaya komputasinya yang sangat tinggi menjadi kendala utama dalam aplikasi dunia nyata. Sebagai solusi jalan tengah yang menyeimbangkan efisiensi *Bi-Encoder* dan efektivitas *Cross-Encoder*, arsitektur *Late Interaction* seperti ColBERT (*Contextualized Late Interaction over BERT*) mulai diterapkan. Khattab dan Zaharia (2020) menunjukkan bahwa dengan menunda interaksi antara kueri dan dokumen hingga tahap akhir representasi token, sistem dapat mempertahankan kecepatan pencarian vektor sekaligus menangkap nuansa semantik yang halus. Pendekatan ini

sangat relevan untuk dokumen hukum yang membutuhkan pencocokan presisi tinggi namun tetap harus melayani pencarian secara *real-time* (Khattab & Zaharia, 2020).

Secara spesifik dalam ranah *Statutory Retrieval* (pencarian pasal undang-undang), tantangan yang dihadapi berbeda dengan pencarian putusan pengadilan (*case law*). Kompetisi internasional COLIEE (*Competition on Legal Information Extraction/Entailment*) secara konsisten menyoroti bahwa pendekatan hibrida sering kali mengungguli model tunggal. Kim et al. (2019) mendemonstrasikan bahwa menggabungkan metode statistik tradisional (TF-IDF) dengan model bahasa neural sangat efektif untuk tugas identifikasi pasal, karena TF-IDF mampu menangkap kata kunci teknis yang spesifik, sementara model neural menangkap konteks semantiknya. Hal ini diperkuat oleh temuan Šavelka dan Ashley (2022) yang menekankan perlunya sistem IR untuk menjembatani kesenjangan makna pada istilah statuta yang sering kali bersifat kaku dan memerlukan konteks eksternal untuk dapat dipahami secara utuh.

Tantangan lain dalam pengembangan IR hukum di lingkungan dengan sumber daya terbatas (*low-resource*), seperti hukum Indonesia, adalah minimnya data berlabel untuk pelatihan (*supervised learning*). Untuk mengatasi hal ini, teknik *Unsupervised Dense Retrieval* dengan pendekatan *Contrastive Learning* menjadi solusi yang menjanjikan. Izacard et al. (2022) membuktikan bahwa model dapat dilatih untuk membedakan dokumen relevan dan tidak relevan tanpa memerlukan ribuan pasangan kueri-dokumen buatan manusia, yang hasilnya bahkan mampu bersaing dengan metode tradisional seperti BM25 pada berbagai *benchmark*. Selain itu, untuk menangani dokumen regulasi yang sangat panjang, pemodelan interaksi pada tingkat paragraf (*Paragraph-Level Interactions*) menggunakan arsitektur seperti BERT-PLI terbukti lebih efektif daripada memproses dokumen secara utuh, karena relevansi hukum sering kali tersembunyi dalam frasa atau ayat spesifik di dalam dokumen yang besar (Shao et al., 2020).

III. Methods

Bab ini menjelaskan material dan metodologi yang digunakan dalam penelitian. Material penelitian mencakup dataset dokumen peraturan keuangan dalam format PDF yang bersumber dari basis data Badan Pemeriksa Keuangan (BPK). Selain itu, bab ini juga menguraikan metode yang diterapkan dalam eksperimen pembangunan sistem *Information Retrieval* (IR), mulai dari pengumpulan data hingga evaluasi.

3.1. Material

Penelitian ini menggunakan dataset yang terdiri dari dokumen peraturan perundang-undangan Republik Indonesia yang relevan dengan domain keuangan. Data tersebut diakuisisi secara sistematis dari basis data resmi Jaringan Dokumentasi dan Informasi Hukum (JDIH) Badan Pemeriksa Keuangan (BPK) yang dapat diakses melalui laman peraturan.bpk.go.id. Pemilihan dokumen dilakukan berdasarkan kategori subjek (subject-based filtering) untuk memastikan bahwa seluruh data yang digunakan memiliki linearitas yang kuat dengan topik keuangan dan regulasi fiskal.

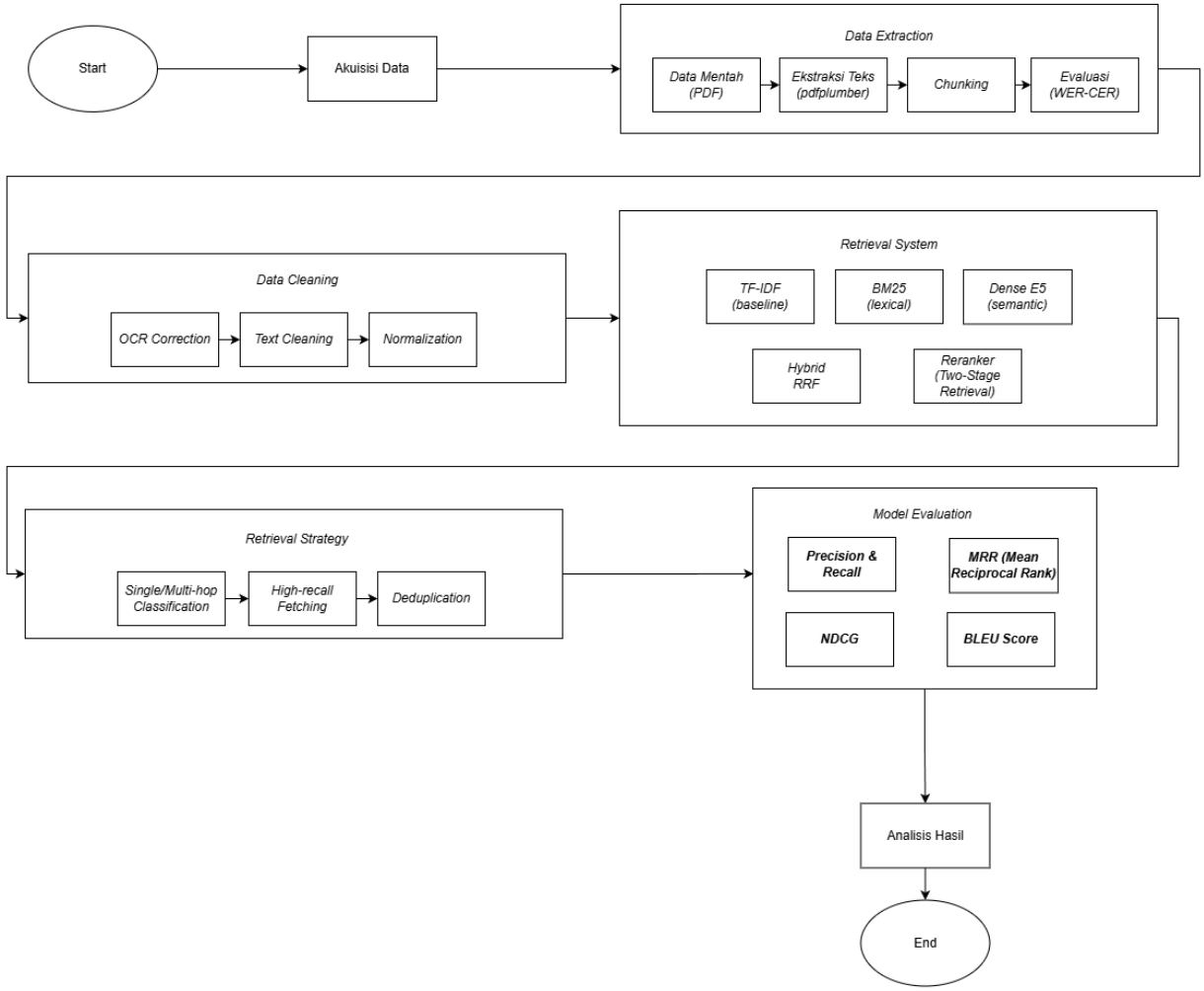
Korpus data ini mencakup sembilan subjek utama yang menjadi fondasi regulasi keuangan negara, yaitu: (1) Keuangan Negara, (2) Anggaran Pendapatan dan Belanja Negara/Daerah (APBN/APBD), (3) Perpajakan, (4) Penerimaan Negara Bukan Pajak (PNBP), (5) Bea dan Cukai, (6) Perbankan dan Sistem Keuangan, (7) BUMN dan Pengelolaan Kekayaan Negara, (8) Pengadaan Barang dan Jasa (PBJ), serta (9) Pemeriksaan atau Audit Keuangan.

Setiap dokumen regulasi dari subjek-subjek tersebut diunduh dalam format Portable Document Format (PDF), yang memuat teks lengkap peraturan beserta pasal-pasal di dalamnya. Untuk menjaga integritas dan ketersediaan data bagi seluruh anggota peneliti, seluruh berkas

PDF yang telah diakuisisi dan disimpan ke dalam repositori penyimpanan terpusat. Dataset ini kemudian melalui tahap pra-pemrosesan (preprocessing) untuk memisahkan teks pasal dari elemen non-tekstual sebelum digunakan dalam eksperimen Information Retrieval.

3.2.Methodology

Penelitian ini menerapkan arsitektur *Single-Stage Retrieval* dengan skema perbandingan *multi-model*. Alur kerja sistem terdiri dari empat tahapan utama: (1) *Data Ingestion*, (2) *Preprocessing*, (3) *Indexing & Retrieval*, dan (4) *Evaluation*, seperti yang ditampilkan pada Gambar 1.



Gambar 1. *Flowchart* Penelitian

Berikut adalah penjelasan rinci dari setiap tahapan yang terlibat dalam eksperimen ini.

3.2.1. Pengumpulan Data & Pra-pemrosesan

Data yang digunakan dalam penelitian berupa dokumen dalam format PDF yang diperoleh melalui teknik *web scraping.* Dokumen-dokumen kemudian di proses untuk menghasilkan *corpus* teks yang siap digunakan dalam sistem *Information Retrieval.*

a. Ekstraksi Teks Dokumen

Dokumen peraturan perundang-undangan dalam format PDF diekstraksi menggunakan *library* pdfplumber untuk memperoleh seluruh isi teks secara terstruktur. pdfplumber dipilih karena kemampuan *layout analysis*-nya yang lebih stabil dalam menangani format dokumen peraturan yang memiliki banyak *whitespace* dan struktur kolom. Proses ekstraksi ini menghasilkan teks mentah dari setiap halaman dokumen, termasuk bagian utama maupun elemen tambahan yang masih tercampur di dalamnya.

b. Pembersihan Teks (*Text Cleaning)*

Tahap pembersihan teks untuk menghilangkan *noise* yang tidak relevan dengan substansi hukum serta menyamakan representasi teks. Proses ini dilakukan dengan menghapus elemen non-informatif seperti nomor halaman, kop surat resmi, bagian pembuka peraturan, tautan URL, dan karakter khusus yang tidak memiliki nilai semantik. Selain itu, seluruh teks dinormalisasi dengan mengonversinya ke huruf kecil (*lowercase)* serta dilakukan standarisasi karakter agar hanya menyisakan huruf, angka, dan tanda baca tertentu. Perbaikan kesalahan OCR ringan juga dilakukan untuk mengatasi karakter yang keliru terbaca. Tahapan ini menghasilkan teks yang konsisten sehingga meningkatkan kualitas pencocokan dokumen pada tahap *retrieval.* Proses pembersihan dilakukan untuk *corpus* dokumen maupun teks *query user.*

c. Segmentasi Dokumen (*Chunking)*

Dokumen peraturan yang bersifat panjang dipecah menjadi bagian-bagian yang lebih kecil agar lebih efektif dalam proses pencarian. Segmentasi dilakukan dengan pendeketan struktur hukum, yaitu memecah dokumen menjadi segmen pasal menggunakan pola *Reguler Expression* yang mengidentifikasi penanda pasal seperti “Pasal 1”, “Pasal 2”, dan seterusnya. Dengan pendekatan ini, sistem *Information Retrieval* dapat melakukan pencarian pada tingkat pasal, bukan pada keseluruhan dokumen, sehingga hasil yang dikembalikan menjadi lebih spesifik dan relevan dengan kebutuhan informasi pengguna.

d. Injeksi Metadata Dokumen

Setiap segmen pasal dilengkapi dengan konteks dokumen melalui proses injeksi metadata. Informasi identitas peraturan, seperti judul atau nomor undang-undang, digabungkan ke dalam teks pasal sehingga setiap segmen tidak hanya memuat substansi hukum, tetapi juga konteks asal dokumennya. Injeksi metadata ini penting untuk menjaga konteks hukum pada saat proses representasi teks, khususnya pada metode pencarian berbasis *semantic embedding,* sehingga pasal-pasal dengan redaksi serupa namun berasal dari peraturan yang berbeda dapat dibedakan secara lebih akurat.

e. Normalisasi dan Kanonisasi Pasal

Untuk menjaga konsistensi struktur data, dilakukan proses normalisasi dan kanonisasi penamaan pasal. Berbagai variasi penulisan pasal yang muncul akibat

perbedaan format dokumen atau hasil ekstraksi teks diseragamkan ke dalam satu format baku. Proses ini memastikan bahwa setiap segmen pasal memiliki identitas yang konsisten, sehingga memudahkan pemetaan antara *corpus* dokumen dan data *ground truth* pada tahap evaluasi sistem.

f. Normalisasi Identitas Dokumen

Tahap akhir pra-pemrosesan mencakup normalisasi identitas dokumen untuk mengatasi variasi penamaan file dan sumber dokumen. Proses ini dilakukan dengan menyamakan identitas dokumen melalui penghapusan ekstensi file, simbol tambahan, serta perbedaan penulisan yang tidak signifikan. Normalisasi identitas dokumen bertujuan untuk memastikan kesesuaian antara *corpus* dokumen dan data *ground truth,* sehingga proses evaluasi sistem *Information Retrieval* dapat dilakukan secara akurat dan adil.

3.2.2. Arsitektur Sistem Pencarian (Retrieval System)

Penelitian ini mengimplementasikan dan membandingkan beberapa pendekatan *Information Retrieval* yang mencakup metode berbasis leksikal, semantik, serta pendekatan gabungan (*hybrid)*. Setiap metode retrieval menghasilkan daftar dokumen atau segmen pasal yang diurutkan berdasarkan skor relevansi terhadap *query user,* yang kemudian dievaluasi menggunakan metrik kinerja tertentu

1. TF-IDF

   Pendekatan TF-IDF digunakan sebagai salah satu metode *baseline* berbasis leksikal. Pada tahap ini, setiap segmen pasal pada *corpus* direpresentasikan dalam bentuk vektor TF-IDF menggunakan *library TfidfVectorizer.* Konfigurasi *word n-gram* digunakan untuk memungkinkan pengenalan frasa hukum yang terdiri dari lebih satu kata. Seluruh teks *corpus* yang telah dinormalisasi diubah ke dalam ruang vektor TF-IDF, sementara setiap *query user* juga ditransformasikan ke dalam representasi vektor yang sama. Skor relevansi dihitung berdasarkan kesamaan antara vektor *query* dan vektor dokumen, kemudian sistem mengembalikan daftar segmen pasal dengan skor tertinggi sebagai hasil pencarian.

2. BM25

   Metode BM25 digunakan sebagai pendekatan *probabilistic retrieval* yang juga berbasis leksikal. *Corpus* yang telah dinormalisasi dan ditokenisasi digunakan untuk membangun model BM25 menggunakan algoritma *BM25Okapi.* Setiap *query* diproses dengan skema tokenisasi yang sama seperti *corpus,* kemudian BM25 menghitung skor relevansi antara *query* dan setiap segmen pasal. Skor tersebut digunakan untuk melakukan perankingan dokumen, dan sejumlah segmen pasal dengan skor tertinggi diambil sebagai hasil *retrieval* awal. Metode ini berperan sebagai *baseline* yang kuat untuk membandingkan efektivitas pendekatan semantik dan *hybrid.*

3. Dense Retrieval

   Pendekatan *dense retrieval* digunakan untuk menangkap kesamaan semantik antara *query* dan dokumen. Pada metode ini, setiap segmen pasal dikodekan menjadi vektor *embedding* menggunakan model *SentenceTransformer* berbasis *multilingual E5*. *Embedding* dokumen disimpan terlebih dahulu untuk mempercepat proses pencarian. Setiap *query user* juga dikodekan menjadi *embedding* dengan skema yang sama,

kemudian skor relevansi dihitung menggunakan *Cosine Similarity* antara vektor kueri v_q dan vektor dokumen v_d sebagai berikut:

$$sim(q,d) = \frac{v_q . v_d}{||v_q|| \cdot ||v_d||}$$

4. Reranking dengan Cross Encoder

   Tahap *reranking* diterapkan untuk meningkatkan kualitas hasil pencarian dengan menilai ulang kandidat dokumen teratas yang diperoleh dari tahap retrieval sebelumnya. Sejumlah kandidat dengan peringkat tertinggi diambil sebagai masukan untuk model *Cross Encoder*. Pada tahap ini, pasangan *query* dan segmen pasal diproses secara bersamaan oleh model *CrossEncoder*, sehingga interaksi antara *query* dan teks dokumen dapat dimodelkan secara lebih mendalam. Model menghasilkan skor relevansi untuk setiap pasangan *query* dokumen, yang kemudian digunakan untuk menyusun ulang peringkat hasil pencarian. Pendekatan ini bertujuan meningkatkan presisi hasil dengan mengorbankan sedikit efisiensi komputasi.

5. Hybrid Retrieval

   Pendekatan *hybrid retrieval* dibangun dengan menggabungkan metode berbasis leksikal dan semantik. Pada penelitian ini, metode *hybrid* mengombinasikan hasil perankingan dari BM25 dan *dense retrieval*. Penggabungan dilakukan menggunakan skema *Reciprocal Rank Fusion (RRF)*, yang mengakumulasi kontribusi peringkat dari masing-masing metode untuk setiap dokumen. Skor fusi dihitung dengan rumus:

$$RFFscore(d) = \sum_{m \in M} \frac{1}{k + r_m(d)}$$

   Dimana:

   - M adalah himpunan metode (BM25 dan Dense E5).
   - $r_m(d)$ adalah peringkat dokumen d pada metode m.
   - k adalah konstanta mitigasi outlier (ditetapkan k=60 sesuai standar literatur).

   Pendekatan ini memungkinkan sistem untuk mempertimbangkan baik kecocokan kata kunci maupun kesamaan semantik dalam menentukan relevansi dokumen. Hasil akhir berupa daftar segmen pasal yang telah difusi dan diurutkan berdasarkan skor gabungan, yang kemudian dapat digunakan langsung atau diteruskan ke tahap *reranking*.

3.2.3. *Evaluasi*

Evaluasi dilakukan menggunakan metode Known-Item Search pada dataset *Ground Truth.* Metrik yang digunakan adalah Recall@K (untuk mengukur kemampuan sistem menemukan dokumen yang benar dalam daftar hasil) dan Mean Reciprocal Rank (MRR) untuk mengukur akurasi peringkat dokumen relevan. Evaluasi dilakukan secara terpisah untuk setiap tema guna memastikan konsistensi performa model pada domain hukum yang berbeda.

IV. Results

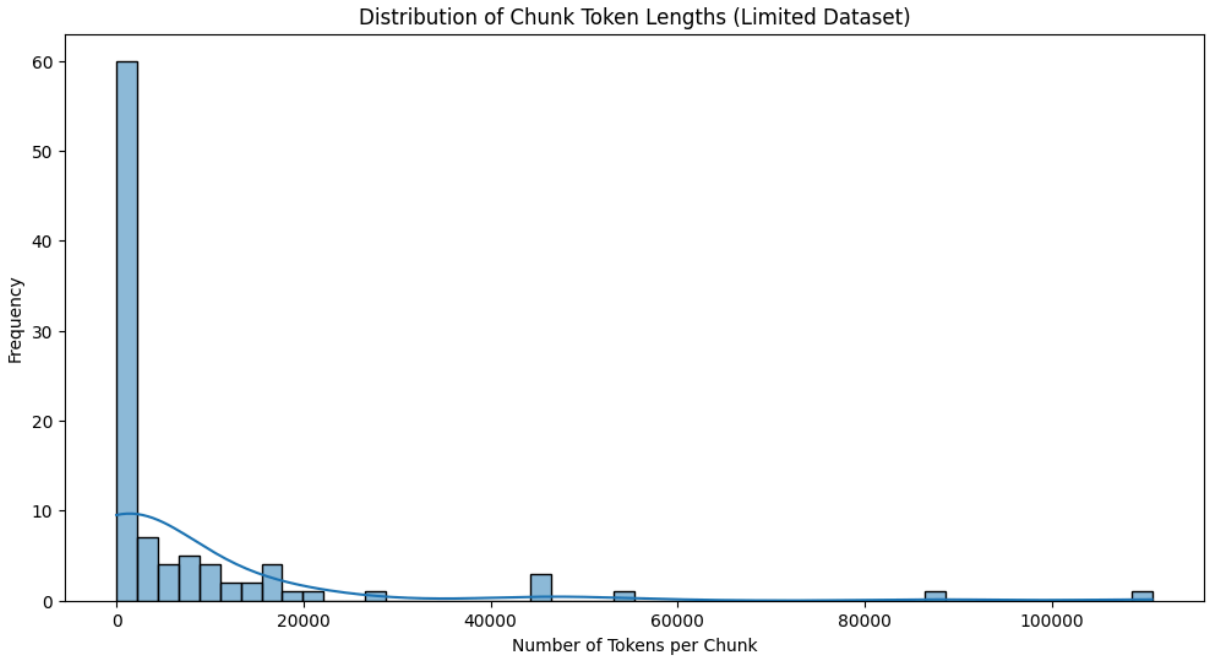
Hasil percobaan yang kami buat disajikan dalam beberapa bagian, dimulai dengan analisis awal berupa *Exploratory Data Analysis* (EDA), yang bertujuan untuk memahami karakteristik data audio yang digunakan, hingga hasil evaluasi model yang diterapkan.

4.1.*Exploratory Data Analysis* (EDA)

Exploratory Data Analysis (EDA) dilakukan untuk menganalisis karakteristik dataset peraturan keuangan, yang meliputi distribusi panjang token (chunk length), analisis frekuensi kata (Word Cloud & Top Words), dan pengecekan kualitas ekstraksi data. Analisis ini bertujuan untuk memastikan data yang digunakan representatif dan valid sebelum masuk ke tahap pemodelan. Berikut adalah hasil EDA yang dilakukan:

4.1.1. Distribusi Panjang Token (Chunk Length)

Analisis panjang token dilakukan untuk memahami karakteristik teks setelah proses segmentasi (chunking). Visualisasi histogram digunakan untuk melihat seberapa panjang potongan teks yang akan diproses oleh model.

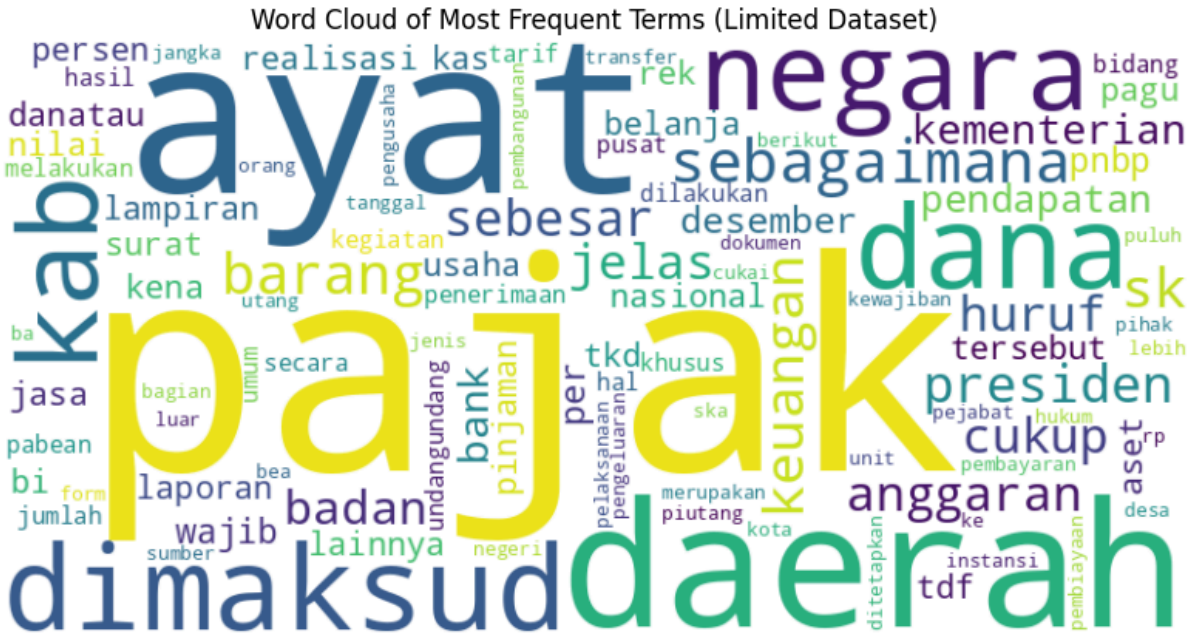


Gambar 2. Visualisasi Distribusi Panjang Token per Chunk

Distribusi panjang token menunjukkan variasi yang sangat ekstrem, dengan panjang token berkisar antara 16 hingga 110.725 kata per chunk, dan nilai rata-rata sekitar 7.281 kata. Grafik menunjukkan pola distribusi yang sangat skewed (miring), di mana terdapat beberapa segmen teks yang sangat panjang (kemungkinan satu UU utuh yang gagal dipecah). Pola ini mengindikasikan bahwa strategi pemecahan teks berbasis paragraf (double newline) saat ini belum optimal untuk dokumen hukum, sehingga diperlukan strategi chunking berbasis Pasal agar panjang input sesuai dengan batasan model Transformer (512 token).

4.1.2. Analisis Frekuensi Kata (Word Cloud & Top Words)

Representasi visual berupa Word Cloud dan tabel frekuensi kata digunakan untuk mengidentifikasi terminologi dominan dan memastikan relevansi konten dengan domain keuangan. Word Cloud & Keyword Dominan Visualisasi di bawah ini merepresentasikan kata-kata yang paling sering muncul dalam korpus, di mana ukuran kata mencerminkan frekuensi kemunculannya.

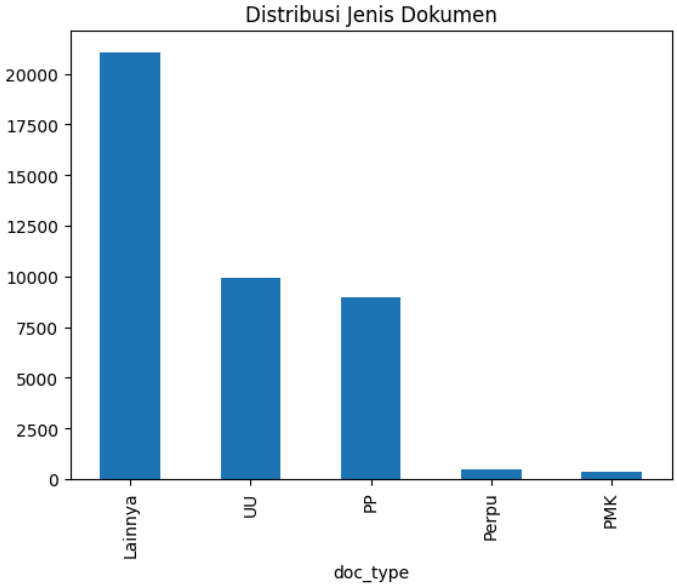


Gambar 3. Word Cloud Dataset Peraturan Keuangan

Berdasarkan analisis frekuensi, kata-kata yang paling dominan adalah "pajak" (6528 kemunculan), diikuti oleh "ayat", "daerah", "dimaksud", dan "dana". Dominasi istilah-istilah ini mengonfirmasi bahwa dataset yang digunakan benar-benar relevan dengan domain regulasi keuangan dan pemerintahan. Keberadaan kata "ayat" dan "dimaksud" yang tinggi juga mencirikan struktur bahasa hukum (legalese) yang khas dalam dokumen peraturan Indonesia.

4.1.3. Distribusi Jenis Dokumen Hukum

Analisis struktur korpus dilakukan melalui visualisasi Diagram Batang (*Bar Chart*) untuk memetakan komposisi dataset berdasarkan hierarki peraturan perundang-undangan. Grafik di bawah ini mengilustrasikan distribusi kuantitatif dari setiap jenis dokumen (seperti Undang-Undang, Peraturan Pemerintah, dan Peraturan Menteri), di mana tinggi batang merepresentasikan total ketersediaan dokumen dalam masing-masing kategori regulasi.



Gambar 4. Bar chart Distrbusi Jenis Dokumen

Berdasarkan visualisasi distribusi jenis dokumen, kategori yang paling dominan adalah Lainnnya, diikuti oleh UU, PP, PERPU, dan PMK. Dominasi ragam regulasi ini mengonfirmasi bahwa dataset yang disusun memiliki cakupan yang komprehensif terhadap berbagai tingkatan hierarki hukum keuangan di Indonesia, mulai dari aturan payung hingga aturan teknis. Keberadaan variasi jenis dokumen dari tingkat Undang-Undang hingga peraturan pelaksana teknis juga mencirikan kompleksitas ekosistem regulasi yang harus ditangani oleh sistem, di mana setiap tingkatan hukum memiliki kedalaman konteks dan bobot otoritas yang berbeda.

4.1.4. Evaluasi Kualitas Ekstraksi Teks (WER & CER)

Sebelum melakukan pemodelan *Information Retrieval*, evaluasi teknis dilakukan untuk mengukur kualitas ekstraksi teks dari format PDF ke format teks terstruktur. Mengingat dokumen hukum memiliki format visual yang kompleks, validasi ini krusial untuk memastikan bahwa noise akibat konversi tidak merusak makna semantik pasal.

Evaluasi dilakukan menggunakan metrik *Word Error Rate* (WER) dan *Character Error Rate* (CER) dengan membandingkan hasil ekstraksi otomatis sistem terhadap *ground truth* manual sebanyak 37 sampel pasal acak. WER mengukur persentase kata yang salah dikenali, disisipkan, atau dihapus, sementara CER mengukur kesalahan pada tingkat karakter yang penting untuk mendeteksi kesalahan ejaan atau typo akibat proses parsing PDF.

Tabel 1. Hasil Evaluasi Kualitas Ekstraksi Teks

Metrik	Mean	Median
WER (Word Error Rate)	17.33%	5.10%
CER (Character Error Rate)	13.28%	2.05%

Hasil evaluasi menunjukkan bahwa sistem ekstraksi memiliki performa yang cukup baik dengan nilai Median WER sebesar 5,10% dan Median CER sebesar 2,05%. Rendahnya nilai median ini mengindikasikan bahwa pada mayoritas dokumen, sistem mampu menyalin teks pasal

dengan tingkat kemiripan yang sangat tinggi terhadap aslinya. Meskipun demikian, nilai Rata-rata (Mean) WER tercatat lebih tinggi pada angka 17,33% dan Mean CER pada 13,28%. Disparitas antara nilai median dan rata-rata (mean) ini menunjukkan adanya sejumlah kecil dokumen outlier yang memiliki format sangat tidak baku atau rusak, sehingga meningkatkan rata-rata kesalahan secara keseluruhan. Namun secara umum, rendahnya median CER (di bawah 3%) menegaskan bahwa sistem mampu menangkap karakter huruf dan angka dalam pasal hukum dengan presisi tinggi, yang merupakan prasyarat vital bagi keberhasilan tahap retrieval selanjutnya.

4.2. Analisis Performa Model

Evaluasi performa model dilakukan untuk membandingkan kemampuan masing-masing pendekatan dalam mengembalikan dokumen/pasal yang relevan terhadap *query* pada domain dokumen hukum Indonesia. Metrik utama yang digunakan adalah precision, recall, MRR, dan nDCG, sedangkan BLEU digunakan sebagai metrik pelengkap untuk melihat kemiripan segmen jawaban top-1 terhadap gold passage. Secara umum, ketika nilai K diperbesar, precision cenderung menurun karena lebih banyak hasil yang diambil, sementara recall meningkat karena peluang menemukan item relevan menjadi lebih tinggi.

4.2.1. Skenario Evaluasi

Evaluasi performa model dilakukan untuk mengukur kemampuan sistem dalam menemukan pasal peraturan perundang-undangan yang relevan terhadap pertanyaan pengguna. Proses evaluasi menggunakan dataset *query* dan *ground truth* yang telah melalui tahap normalisasi teks untuk memastikan konsistensi data.

Pada setiap *query,* sistem menerapkan strategi *retrieval bertahap*, yaitu dengan mengambil hingga 100 kandidat awal dari masing-masing model. Selanjutnya dilakukan proses deduplikasi untuk menghilangkan hasil yang berasal dari dokumen yang sama, sehingga diperoleh maksimal 30 pasal unik terbaik. Evaluasi performa kemudian dilakukan pada tiga tingkat pengambilan hasil teratas, yaitu k = 5, k = 15, dan k = 30. Pendekatan ini digunakan untuk merepresentasikan skenario penggunaan sistem pencarian hukum, di mana pengguna umumnya hanya memperhatikan sejumlah hasil teratas, namun tetap mengharapkan seluruh pasal relevan dapat ditemukan.

4.2.2. Analisis Kinerja Utama pada Peringkat Teratas (Top-5)

Evaluasi pada K=5 merepresentasikan kondisi kritis di mana pengguna hanya melihat halaman pertama hasil pencarian. Metrik MRR dan nDCG sangat vital pada tahap ini untuk menilai kemampuan model dalam menempatkan dokumen paling relevan di posisi teratas.. Tabel 2 menunjukkan performa model pada tahap krusial ini.

Tabel 2. Matriks Evaluasi Model pada Kedalaman k=5

Model	Recall@5	MRR@5	nDCG@5	BLEU Score
HYBRID_RRF	0.477164	0.689778	0.531610	29.65
BM25	0.429386	0.66889	0.496616	28.93
RERANKER_BGE	0.407037	0.602222	0.455364	28.099
DENSE_E5	0.392831	0.593778	0.451754	25.334
TFIDF_WORD	0.381275	0.567	0.445950	19.89

Berdasarkan Tabel 2, model HYBRID_RRF menunjukkan dominasi performa dengan nilai tertinggi di seluruh metrik. Nilai MRR sebesar 0.6898 mengindikasikan bahwa jawaban

benar rata-rata muncul pada posisi 1 atau 2. Temuan ini membuktikan bahwa penggabungan sinyal leksikal (kata kunci) dan semantik (makna) mampu meningkatkan kualitas peringkat secara signifikan dibandingkan model tunggal.

Menariknya, BM25 menempati posisi kedua, mengungguli model berbasis *Deep Learning* murni seperti RERANKER_BGE dan DENSE_E5. Hal ini menunjukkan bahwa dalam dokumen hukum yang kaya istilah teknis, pencocokan kata kunci yang tepat masih memegang peranan vital.

4.2.3. Analisis Skalabilitas pada Kedalaman Menengah (Evaluasi pada K=15)

Pada K=15, fokus evaluasi bergeser untuk melihat apakah model mampu menjaga relevansi ketika pengguna menelusuri hasil lebih jauh.

Tabel 3. Matriks Evaluasi Model pada Kedalaman k=15

Model	Recall@15	MRR@15	nDCG@15	BLEU Score
HYBRID_RRF	0.5337	0.6963	0.5657	29.65
BM25	0.5221	0.6780	0.5449	28.93
RERANKER_BGE	0.5043	0.6117	0.5141	28.099
DENSE_E5	0.4843	0.6083	0.5030	25.33
TFIDF_WORD	0.4549	0.5830	0.4929	19.89

Hasil pada Tabel 3 menunjukkan tren positif pada model HYBRID_RRF, di mana nilai Recall meningkat menjadi 0.5337 dan nDCG naik menjadi 0.5657. Peningkatan nDCG ini mengindikasikan bahwa model *Hybrid* tidak hanya menambah jumlah dokumen relevan, tetapi dokumen-dokumen tambahan tersebut juga memiliki bobot relevansi yang tinggi. Posisi kedua tetap ditempati oleh BM25, menegaskan kestabilan metode leksikal probabilitas dalam menangani kueri hukum.

4.2.4 Analisis Stabilitas pada Kedalaman Maksimum (Evaluasi pada K=30)

Evaluasi pada K=30 mensimulasikan tugas *legal research* yang komprehensif, di mana pengguna menelusuri banyak dokumen untuk memastikan tidak ada pasal yang terlewat.

Tabel 4. Hasil Evaluasi Model pada K=30

Model	Recall@30	MRR@30	nDCG@30	BLEU Score
HYBRID_RRF	0.5784	0.6983	0.5830	29.65
BM25	0.5670	0.6799	0.5608	28.93
RERANKER_BGE	0.5532	0.6137	0.5285	28.099
DENSE_E5	0.5277	0.6109	0.5251	25.33
TFIDF_WORD	0.5038	0.5862	0.5141	19.89

Pada tahap evaluasi terdalam ini (Tabel 4), HYBRID_RRF tetap konsisten memimpin dengan Recall sebesar 0.5784. Stabilitas nilai MRR yang bertahan di angka ~0.69 menunjukkan bahwa penambahan jumlah dokumen (K) tidak mendegradasi kualitas peringkat hasil teratas. Model BM25 terus membayangi performa Hybrid dengan selisih yang tipis, sementara TFIDF_WORD konsisten berada di posisi terbawah, menandakan keterbatasan metode statistik sederhana dalam menangkap konteks kompleks dokumen hukum.

4.2.5. Analisis Kualitas Teks Jawaban (BLEU Score)

Selain akurasi peringkat, penelitian ini juga mengevaluasi kemiripan tekstual antara segmen jawaban teratas (Top-1*)* yang diberikan model dengan teks pada *gold passage* menggunakan metrik BLEU Score. Analisis ini dipisahkan untuk melihat apakah dokumen yang dianggap relevan secara sistem juga memiliki konten teks yang presisi sesuai rujukan.

Tabel 5. Perbandingan BLEU Score Antar Model

Model	BLEU Score
HYBRID_RRF	29.65
BM25	28.93
RERANKER_BGE	28.099
DENSE_E5	25.33
TFIDF_WORD	19.89

Berdasarkan Tabel 5, model HYBRID_RRF mencatatkan skor BLEU tertinggi (29.65), diikuti dengan sangat ketat oleh BM25 (28.93).

- Implikasi: Tingginya skor BLEU pada model yang melibatkan komponen leksikal (Hybrid dan BM25) menunjukkan bahwa model-model ini sangat efektif dalam mengembalikan pasal dengan frasa yang *exact* atau sangat mirip dengan kunci jawaban. Ini sangat krusial dalam hukum di mana perbedaan satu kata dapat mengubah makna pasal.
- Perbandingan: Model semantik murni (DENSE_E5) memiliki skor BLEU yang lebih rendah (25.33). Hal ini wajar karena model semantik cenderung menangkap "makna" atau parafrasa, sehingga dokumen yang dikembalikan mungkin relevan secara topik tetapi memiliki struktur kalimat yang berbeda dari *ground truth*, mengakibatkan skor BLEU yang lebih rendah.
- Kelemahan TF-IDF: Skor terendah pada TFIDF_WORD (19.89) mengindikasikan bahwa meskipun model ini menemukan kata kunci yang sama, segmen teks yang diambil sering kali kurang akurat atau terpotong.

4.2.6. Ringkasan Analisis Performa

Berdasarkan keseluruhan hasil evaluasi, dapat disimpulkan bahwa pendekatan hybrid yang dikombinasikan dengan reranker memberikan performa terbaik dalam sistem pencarian pasal peraturan perundang-undangan. Pendekatan ini tidak hanya mampu meningkatkan cakupan pasal relevan, tetapi juga menjaga kualitas urutan hasil pencarian yang ditampilkan kepada pengguna. Dengan demikian, kombinasi model leksikal, semantik, dan reranking merupakan strategi yang paling efektif untuk diterapkan dalam sistem pencarian hukum berbasis teks.

V. Discussion

Berdasarkan hasil eksperimen yang telah dilakukan, dapat disimpulkan bahwa pencarian pasal hukum dalam domain regulasi keuangan memiliki kompleksitas yang tinggi, baik dari sisi karakteristik bahasa maupun struktur dokumen. Hasil evaluasi menunjukkan bahwa pendekatan berbasis leksikal seperti BM25 dan TF-IDF masih mampu memberikan performa yang stabil pada kondisi tertentu, terutama ketika istilah dalam pertanyaan memiliki kesesuaian langsung dengan terminologi hukum dalam dokumen. Namun demikian, performa model leksikal cenderung menurun ketika dihadapkan pada variasi bahasa alami pengguna yang lebih informal,

sehingga menguatkan temuan sebelumnya terkait permasalahan *vocabulary mismatch* pada sistem information retrieval hukum.

Model berbasis semantik, khususnya Dense Retrieval menggunakan embedding E5, menunjukkan peningkatan yang konsisten pada metrik recall. Hal ini mengindikasikan bahwa representasi semantik mampu menangkap hubungan makna antara pertanyaan dan pasal hukum meskipun tidak terdapat kecocokan kata secara eksplisit. Temuan ini sejalan dengan literatur yang menyatakan bahwa dense retrieval unggul dalam memahami makna implisit, terutama pada domain khusus seperti hukum. Namun, meskipun recall meningkat, model dense masih menghadapi tantangan dalam menempatkan dokumen paling relevan pada peringkat teratas, yang tercermin dari nilai MRR dan nDCG yang tidak selalu unggul dibandingkan pendekatan lain.

Pendekatan hybrid yang mengombinasikan sinyal leksikal dan semantik terbukti memberikan performa yang lebih seimbang. Hybrid RRF mampu mempertahankan recall tinggi sekaligus memperbaiki stabilitas peringkat dibandingkan penggunaan model tunggal. Temuan ini menguatkan hasil penelitian sebelumnya yang menyatakan bahwa penggabungan metode statistik dan neural merupakan strategi efektif dalam statutory retrieval. Hal ini menunjukkan bahwa pada dokumen hukum, informasi relevan sering kali bergantung pada kombinasi kecocokan istilah teknis dan pemahaman konteks semantik.

Penggunaan reranker berbasis cross-encoder memberikan peningkatan signifikan pada kualitas hasil teratas, terutama pada metrik MRR dan nDCG. Model ini mampu memproses interaksi mendalam antara pertanyaan dan teks pasal, sehingga lebih akurat dalam menentukan urutan relevansi. Namun demikian, peningkatan akurasi ini datang dengan konsekuensi biaya komputasi yang jauh lebih tinggi dibandingkan model lain. Kondisi ini menjadi keterbatasan penting apabila sistem diterapkan pada skenario dunia nyata yang menuntut efisiensi waktu dan sumber daya, khususnya pada korpus hukum berskala besar.

Analisis BLEU score menunjukkan perbedaan kualitas teks hasil pencarian yang cukup mencolok antara model statistik dan model neural. Model neural menghasilkan teks pasal yang secara struktural dan semantik lebih mendekati jawaban rujukan, sedangkan model leksikal cenderung mengembalikan pasal dengan relevansi kata kunci namun kurang tepat secara konteks. Temuan ini mengindikasikan bahwa evaluasi berbasis kesamaan teks menjadi pelengkap penting dalam menilai kualitas sistem retrieval hukum, terutama ketika sistem digunakan sebagai pendukung pengambilan keputusan.

Meskipun sistem yang dikembangkan menunjukkan performa yang menjanjikan, terdapat beberapa keterbatasan yang perlu dicermati. Pertama, kualitas hasil retrieval masih sangat bergantung pada keberhasilan proses ekstraksi dan segmentasi pasal. Kesalahan chunking atau pasal yang terlalu panjang dapat menurunkan efektivitas model Transformer yang memiliki batasan panjang input. Kedua, pendekatan evaluasi masih berfokus pada retrieval pasal, sehingga aspek penalaran hukum lintas pasal (*multi-hop reasoning*) belum sepenuhnya dimodelkan secara eksplisit. Hal ini terlihat dari masih terbatasnya kemampuan sistem dalam menangani pertanyaan yang memerlukan integrasi informasi dari beberapa pasal sekaligus.

Untuk mengatasi keterbatasan tersebut, penelitian selanjutnya dapat mempertimbangkan penggunaan arsitektur *late interaction* seperti ColBERT yang menawarkan keseimbangan antara efisiensi dan akurasi, atau integrasi sistem retrieval dengan modul Question Answering berbasis

Large Language Models yang memiliki kemampuan penalaran lebih kompleks. Selain itu, penerapan teknik *Automatic Query Expansion* dan *domain-adaptive pretraining* pada korpus hukum Indonesia berpotensi meningkatkan performa sistem dalam menghadapi variasi bahasa pengguna yang lebih luas. Dengan demikian, sistem statutory retrieval dapat dikembangkan menjadi lebih robust, efisien, dan relevan untuk mendukung pencarian pasal hukum di Indonesia.

VI. Conclusion

Eksperimen menunjukkan bahwa pencarian pasal pada regulasi keuangan Indonesia menantang karena perbedaan bahasa sehari-hari dan terminologi hukum serta struktur dokumen PDF yang kompleks. Metode hybrid yang menggabungkan model lexical dan semantik menggunakan skema *Reciprocal Rank Fusion (RRF)* merupakan pendekatan terbaik dan paling konsisten pada seluruh tingkat K karena mampu menangkap lebih banyak hasil relevan sekaligus menjaga kualitas urutan peringkat. BM25 menjadi baseline yang kuat dan stabil, menegaskan pentingnya pencocokan kata kunci pada dokumen hukum. Dense Retrieval membantu menangkap kemiripan semantik tetapi belum dapat melampaui hybrid atau BM25, sedangkan TF-IDF memberikan performa terendah. Tahap reranking (BGE) belum menunjukkan peningkatan dominan pada MRR dan nDCG dibanding hybrid pada konfigurasi eksperimen ini.

Untuk pengembangan selanjutnya, sistem dapat ditingkatkan dengan memperkuat tahap ekstraksi dan segmentasi dokumen agar sesuai batas token model dan mengurangi error dari PDF yang menjadi outlier. Di sisi pencarian, penerapan *automatic query expansion* dan normalisasi istilah berpotensi menekan *vocabulary mismatch*, serta pelatihan adaptif domain pada korpus hukum Indonesia dapat meningkatkan kualitas embedding dan reranking. Selain itu, eksplorasi arsitektur retrieval yang lebih efisien penting untuk menyeimbangkan akurasi dan biaya komputasi, sekaligus menambahkan mekanisme *multi-hop reasoning* untuk pertanyaan lintas pasal. Terakhir, evaluasi perlu diperluas dengan *ground truth* yang lebih beragam dan skenario pengguna nyata agar kinerja sistem lebih representatif saat diterapkan di dunia nyata.

References

Kaur, H., & Kaur, P. (2019). *Text Mining: Techniques, Applications and Issues*.

Shoukat Ali, A. A., & Shandilya, V. K. (2021). *AI–Natural Language Processing (NLP)*. International Journal for Research in Applied Science & Engineering Technology.

Taylor, M. (2019). *AI and Law Librarians: Introducing the Idea of Creating a Legal Information Research Team to Prepare Students for the Practice of Law*.

Hong, L., et al. (n.d.). *Advanced Text Documents Information Retrieval System for Search Services*.

Bonifacio, L., Jeronymo, R., Abonizio, H. Q., Campiotti, I., Fadaei, M., Lotufo, R., & Nogueira, R. (2021). mMARCO: A Multilingual Version of the MS MARCO Passage Ranking Dataset. arXiv preprint arXiv:2108.13897.

Chalkidis, I., Jana, A., Hartung, D., Bommarito, M., Androutsopoulos, I., Katz, D. M., & Aletras, N. (2021). LexGLUE: A Benchmark Dataset for Legal Language Understanding in English. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL), 4310–4330.

Curran, M., O'Sullivan, D., & Lewis, D. (2023). Legal Information Retrieval with Large Language Models: A Comprehensive Survey. Artificial Intelligence and Law, 31, 1–35.

Louis, A., van Dijck, G., & Spanakis, G. (2022). A Statutory Article Retrieval Dataset in French. Proceedings of the Natural Legal Language Processing Workshop 2022, 156–166.

Pradeep, R., Ma, X., & Lin, J. (2021). The Expando-Mono-Duo Design Pattern for Text Ranking with Pretrained Sequence-to-Sequence Models. arXiv preprint arXiv:2101.05667.

Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., & Gurevych, I. (2021). BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models. Proceedings of the Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track.

Adhikari, N. S., & Agarwal, S. (2025). A Comparative Study of PDF Parsing Tools Across Diverse Document Categories. *arXiv preprint arXiv:2410.09871*.

Aprilio, P., Michael, Nugraha, P. S., & Fahmi, H. (2025). Hybrid Feature Combination of TF-IDF and BERT for Enhanced Information Retrieval Accuracy. *JISA (Jurnal Informatika dan Sains)*, 8(1).

Ariai, F., Mackenzie, J., & Demartini, G. (2024). Natural Language Processing for the Legal Domain: A Survey of Tasks, Datasets, Models, and Challenges.

Chang, W.-C., Yu, F. X., Chang, Y.-W., Yang, Y., & Kumar, S. (2020). Pre-training Tasks for Embedding-based Large-scale Retrieval. In *International Conference on Learning Representations (ICLR)*.

Chugh, A., Sharma, V. K., Kumar, S., Nayyar, A., Qureshi, B., Bhatia, M. K., & Jain, C. (2021). Spider Monkey Crow Optimization Algorithm With Deep Learning for Sentiment Classification and Information Retrieval. *IEEE Access*, *9*, 24262–24279.

Fang, Y., Mao, J., Zhan, J., Su, W., & Liu, Y. (2023). Scaling Laws For Dense Retrieval. *arXiv preprint*.

Inje, B., Nagwanshi, K. K., & Rambola, R. K. (2024). An efficient document information retrieval using hybrid global search optimization algorithm with density based clustering technique. *Cluster Computing*, *27*, 689–705.

Jiang, S., & Li, Q. (2023). Research and Implementation of PDF Specific Element Fast Extraction. *2023 4th International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE)*.

Kulkarni, M., & Kale, V (2021). Information Retrieval based Improvising Search using Automatic Query Expansion. *Proceedings of the Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV 2021)*.

Keller, J., & Munz, L. P. M. (2021). TEKMA at CLEF-2021: BM25 based rankings for scientific publication retrieval and data set recommendation. In *Proceedings of the Working Notes of CLEF 2021*.

Li, S. (2024). A Cross Language Information Retrieval Model Based on Latent Semantic Analysis. *Intelligent Computing Technology and Automation*, IOS Press.

Marwah, D., & Beel, J. (2021). Term-Recency for TF-IDF, BM25 and USE Term Weighting. *School of Computer Science and Statistics, Trinity College Dublin*.

Rezvani, S., Naghshineh, N., & Khalilijafarabad, A. (2023). Implementation of Experts' Retrieval Model Using Latent Semantic Indexing (LSA) Method and Temporal Graph. *Library and Information Science Research*, 13(1), 226–245.

Izacard, G., Caron, M., Hosseini, L., Riedel, S., Bojanowski, P., Joulin, A., & Grave, E. (2022). Unsupervised Dense Information Retrieval with Contrastive Learning. *Transactions on Machine Learning Research*.

Khattab, O., & Zaharia, M. (2020). ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 39–48.

Kim, M.-Y., Rabelo, J., & Goebel, R. (2019). Statute Law Information Retrieval and Entailment. *Proceedings of the Competition on Legal Information Extraction/Entailment (COLIEE)*.

Kim, S., Park, H., & Lee, J. (2020). Word2vec-based latent semantic analysis (W2V-LSA) for topic modeling: A study on blockchain technology trend analysis. *Expert Systems with Applications*, *152*, 113401.

Souza, E., Vitório, D., Moriyama, G., Santos, L., Martins, L., Souza, M., Fonseca, M., Félix, N., Carvalho, A. C. P. L. F., Albuquerque, H. O., & Oliveira, A. L. I. (2021). *An Information Retrieval Pipeline for Legislative Documents from the Brazilian Chamber of Deputies*. https://doi.org/10.3233/FAIA210326

Shao, Y., Mao, J., Liu, Y., Ma, W., Satoh, K., Zhang, M., & Ma, S. (2020). BERT-PLI: Modeling Paragraph-Level Interactions for Legal Case Retrieval. *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20)*, 3501–3507.

Šavelka, J., & Ashley, K. D. (2022). Legal Information Retrieval for Understanding Statutory Terms. *Artificial Intelligence and Law*, 30, 245–289.

Verma, J. P., Bhargav, S., Bhavsar, M., Bhattacharya, P., Bostani, A., Chowdhury, S., Webber, J., & Mehbodniya, A. (2023). Graph-Based Extractive Text Summarization Sentence Scoring Scheme for Big Data Applications. *Information*, *14*(9), 472. https://doi.org/10.3390/info14090472

Lampiran 1. Drive Scraping

Link :
https://drive.google.com/drive/folders/1wX5BwMJkabNvp0W5HvGLlf4DPoj-Xyjs?usp=sharing