# BAB I
# PENDAHULUAN

## 1.1 Latar Belakang

Pesatnya kemajuan teknologi informasi telah mengubah cara penyimpanan dan pengelolaan data secara mendasar, hal ini memicu lonjakan data yang luar biasa di berbagai sektor. Proyeksi *Global DataSphere* mencatat pertumbuhan data digital mencapai *Compound Annual Growth Rate* (CAGR) sekitar 23% per tahun [1]. Di lingkup nasional, keterbukaan informasi peradilan Indonesia telah menghasilkan lebih dari 10,5 juta dokumen putusan dalam Direktori Putusan Mahkamah Agung RI hingga akhir April 2026, dengan rata-rata lebih dari 800 ribu dokumen baru diunggah setiap tahunnya [2]. Melimpahnya jumlah dokumen ini menghadirkan tantangan nyata bagi hakim yang memerlukan referensi preseden, pengacara yang membutuhkan dukungan argumen hukum, serta akademisi dan mahasiswa hukum yang memerlukan dokumen acuan berdasarkan kemiripan kasus.

Urgensi tantangan ini diakui oleh institusi peradilan tinggi di Indonesia. Dalam sidang istimewa Laporan Tahunan MA tahun 2023, ketua Mahkamah Agung Dr. H. M. Syarifuddin, S.H., M.H., secara terbuka menyatakan bahwa tidak jarang ditemukan putusan pengadilan yang saling bertentangan untuk perkara yang serupa [3]. Di sisi lain, jumlah putusan pengadilan yang mencapai 800 ribu per tahun menjadikan proses pencarian manual menjadi tidak efisien [4]. Selain itu, Sekretaris Jenderal Mahkamah Konstitusi, Heru Setiawan, menyatakan bahwa besarnya volume perkara dan data hukum menyebabkan hakim memerlukan dukungan sistem informasi yang mampu mempermudah akses, pencarian, dan analisis putusan serta argumen hukum secara efisien guna mendukung konsistensi dan kualitas putusan [5].

Sistem pencarian pada Direktori Putusan MA RI saat ini masih berbasis kata kunci (*keyword-based)*. Sistem pencarian berbasis kata kunci memiliki keterbatasan mendasar akibat fenomena *vocabulary mismatch*, yaitu perbedaan penggunaan istilah yang merujuk pada topik yang sama serta ketidakmampuan menangkap kesamaan makna pada dokumen panjang dan tidak terstruktur [6]. Kondisi ini

menunjukkan bahwa sistem yang ada masih memiliki ruang untuk ditingkatkan dalam mendukung identifikasi kasus yang memiliki kemiripan substansial.

Dalam konteks *Information Retrieval* (IR), pencarian merujuk pada proses menemukan dokumen atau informasi yang relevan dari kumpulan data besar berdasarkan kebutuhan informasi pengguna, sedangkan pencocokan merupakan mekanisme komputasional untuk membandingkan representasi *query* dengan representasi dokumen guna menghitung tingkat kemiripan atau relevansi. Manning et al. menjelaskan bahwa IR berfokus pada penemuan dokumen teks tidak terstruktur yang memenuhi kebutuhan informasi pengguna [7], sementara Jiang dan Cai menyatakan bahwa *text matching* berperan penting dalam tugas IR karena digunakan untuk mengukur kesesuaian antara teks atau dokumen [8]. Oleh karena itu, penelitian ini menggunakan pendekatan pencocokan karena tujuan utamanya bukan sekadar menemukan dokumen berdasarkan kata kunci, melainkan membandingkan kasus baru dengan dokumen putusan terdahulu untuk memperoleh dokumen yang memiliki kemiripan substansial, baik dari aspek semantik maupun struktur fakta hukum. Dengan demikian, pencarian menjadi tujuan akhir sistem, sedangkan pencocokan menjadi proses utama untuk menentukan dokumen yang paling relevan berdasarkan representasi semantik dan representasi struktural [7], [8].

Untuk memahami mengapa sistem yang ada belum memadai, perlu dipahami evolusi teknik *Information Retrieval* (IR). IR mengandalkan teknik pencocokan dokumen sebagai mekanisme intinya, yang bertujuan mengukur kemiripan atau relevansi antar dokumen berdasarkan representasi tertentu [8]. Menurut Jiang *et al.* [9], perkembangan metode *retrieval* berkembang dari pendekatan berbasis *string* dan statistik korpus menuju pendekatan jaringan saraf dan metode berbasis struktur graf. Perkembangan *Natural Language Processing* (NLP) kemudian mendorong pergeseran menuju representasi semantik berbasis transformer, di mana arsitektur BERT terbukti lebih efektif dalam menangkap nuansa semantik dibandingkan metode konvensional seperti TF-IDF [10].

Namun, mengandalkan representasi semantik saja belum memadai untuk domain hukum. Meskipun model transformer mampu menangkap makna kata secara efektif, model ini masih memiliki keterbatasan dalam menangkap

pengetahuan faktual dan hubungan terstruktur antar entitas secara eksplisit [11]. Penelitian sebelumnya menunjukkan bahwa pendekatan berbasis semantik maupun leksikal masih belum mampu memahami relevansi kasus hukum secara komprehensif [12]. Dalam konteks dokumen hukum, kelemahan ini memiliki konsekuensi konkret. Sebagai contoh, kalimat "Terdakwa mengambil barang milik orang lain tanpa izin" dan "Terdakwa menggunakan barang milik orang lain tanpa izin" memiliki kemiripan kosakata dan konteks linguistik sehingga pendekatan berbasis semantik berpotensi menganggap keduanya relevan. Namun secara hukum, keduanya merepresentasikan tindak pidana yang berbeda pencurian dan penggelapan dengan implikasi hukum yang tidak sama [13]. Keterbatasan ini dikonfirmasi secara empiris oleh penelitian Tang et al. [14] yang secara eksplisit menyatakan bahwa model berbasis *language model* mengabaikan informasi struktural hukum yang penting seperti hubungan antar para pihak, tindak pidana, dan bukti karena hanya mengandalkan teks mentah.

Komunitas riset internasional mulai menyadari bahwa pendekatan berbasis teks saja belum mampu memahami relevansi kasus hukum secara komprehensif. Ma et al. [12] menunjukkan bahwa integrasi informasi struktural dapat meningkatkan performa *legal case retrieval*, sementara Tang et al. [14] melalui CaseGNN membuktikan bahwa representasi graf dokumen hukum mampu mengungguli metode *state of the art* pada benchmark COLIEE. Pan et al. [11] juga menunjukkan bahwa *Knowledge Graph* dapat melengkapi keterbatasan LLM dalam merepresentasikan relasi antar entitas secara eksplisit. Namun, penelitian tersebut masih berfokus pada korpus hukum berbahasa Inggris dan belum dirancang untuk konteks hukum Indonesia. Di sisi lain, penelitian NLP hukum Indonesia masih terbatas [21] dan belum mengintegrasikan transformer, *Knowledge Graph* berbasis *triples* (SPO), dan *Case-Based Reasoning* dalam satu sistem legal *retrieval* terpadu. penelitian terdahulu umumnya hanya menggunakan salah satu komponen, misalnya pendekatan semantik saja, struktur graf saja, atau CBR saja, sedangkan penelitian ini mengintegrasikan ketiganya dalam satu sistem retrieval.

*Knowledge Graph* (KG) berbasis *triples* Subjek, Predikat, Objek (SPO) menjadi pendekatan yang dapat digunakan untuk merepresentasikan relasi antar entitas secara eksplisit, karena mampu menangkap hubungan kausal yang tidak

selalu terlihat dari permukaan teks [11]. Yi et al. [10] melalui model SKIE menunjukkan bahwa integrasi pengetahuan struktural berbasis graf ke dalam proses *pre-training* meningkatkan performa ekstraksi informasi dibandingkan pendekatan yang hanya mengandalkan teks.

Setelah data direpresentasikan melalui *Knowledge Graph*, tantangan berikutnya adalah mekanisme pencarian yang efektif. Berbeda dengan CaseGNN yang menggunakan *Graph Neural Network* untuk pembelajaran representasi graf secara *end-to-end* [14], penelitian ini mengadopsi *Case-Based Reasoning* (CBR) sebagai kerangka *retrieval*. CBR adalah paradigma pemecahan masalah yang menyelesaikan masalah baru dengan mengadaptasi solusi dari kasus-kasus serupa yang pernah ada sebelumnya, sehingga secara konseptual selaras dengan cara hakim mencari preseden yaitu membandingkan kasus baru dengan basis kasus terdahulu yang paling mirip. Wiratunga et al. [15] dalam studi CBR-RAG membuktikan bahwa penggunaan CBR untuk mengambil kasus serupa berkontribusi pada peningkatan kualitas jawaban dibandingkan sistem tanpa CBR pada tugas *legal question answering*.

Berdasarkan uraian di atas, penelitian ini mengusulkan pengembangan sistem *Structural Semantic Retrieval* yang mengintegrasikan tiga komponen utama yakni Indo-LegalBERT sebagai *encoder* semantik yang sudah di *fine-tuning* pada domain hukum Indonesia, *Knowledge Graph* berbasis *triples* SPO dengan Node2Vec sebagai representasi relasi antar entitas hukum, dan *Case-Based Reasoning* (CBR) sebagai pencocokan kasus. Integrasi ketiga komponen tersebut dilakukan karena masing-masing memiliki peran yang saling melengkapi. Dokumen putusan pidana umum dipilih sebagai domain uji karena kompleksitas relasi kausal antar entitasnya yang tinggi, serta belum ditemukannya penelitian *legal case retrieval* berbasis integrasi semantik struktural pada domain dan bahasa ini. Kontribusi utama penelitian ini adalah menyediakan bukti empiris apakah integrasi representasi struktural berbasis KG dapat meningkatkan akurasi *retrieval* dokumen hukum Indonesia dibandingkan pendekatan semantik, diukur menggunakan metrik *Precision*@K, *Recall*@K, dan MRR pada dataset Indo-Law.

## 1.2 Rumusan Masalah

Berdasarkan latar belakang masalah yang telah diuraikan, maka rumusan masalah dalam penelitian ini adalah sebagai berikut:

1. Bagaimana merancang arsitektur *Structural Semantic Retrieval* yang mengintegrasikan representasi semantik berbasis Indo-LegalBERT dan representasi struktural berbasis *Knowledge Graph* dengan *Triples* S-P-O (Subjek, Predikat, Objek) untuk merepresentasikan relasi antar entitas dalam dokumen putusan pidana umum Indonesia?
2. Bagaimana kinerja sistem *Structural Semantic Retrieval* yang diusulkan dalam menemukan dokumen hukum yang relevan berdasarkan evaluasi menggunakan metrik *Precision@3*, *Recall@3*, dan *MRR*?

## 1.3 Tujuan Penelitian

Berdasarkan rumusan masalah yang telah dipaparkan, tujuan utama dari penelitian ini adalah:

1. Mengembangkan arsitektur *Structural Semantic Retrieval* yang mengintegrasikan representasi semantik berbasis Indo-LegalBERT dan representasi struktural berbasis *Knowledge Graph* dalam bentuk *triples* (SPO), dalam kerangka *Case-Based Reasoning* (CBR), untuk merepresentasikan hubungan antar entitas pada dokumen putusan pidana umum Indonesia.
2. Mengevaluasi kinerja sistem *Structural Semantic Retrieval* yang diusulkan menggunakan metrik *Precision@K*, *Recall@K*, dan *Mean Reciprocal Rank* (MRR) guna mengetahui kemampuan sistem dalam menemukan dokumen hukum yang relevan menggunakan metrik *Precision@K*, *Recall@K*, dan *MRR*.

## 1.4 Batasan Masalah

Penelitian ini memiliki sejumlah batasan agar ruang lingkup penelitian tetap terarah sesuai dengan tujuan, yaitu:

1. Sumber data dalam penelitian ini terbatas pada *Dataset* Indo-Law yang berisi dokumen putusan pidana umum dalam format XML yang bersumber dari Direktori Putusan Mahkamah Agung RI [16]. Pemrosesan data difokuskan pada tag anotasi <fakta> atau <fakta_hukum> yang memuat

narasi kronologis peristiwa dan pembuktian, sedangkan bagian lain seperti <identitas> dan <amar_putusan> hanya digunakan sebagai metadata pelengkap.

2. Representasi ontologi dalam penelitian ini menggunakan pendekatan *lightweight ontology* berbasis *triples* (SPO) tanpa penerapan metode *graph learning* atau *reasoning* yang kompleks.
3. Tahapan Siklus CBR hanya menggunakan tahap *Retrieve* dan *Reuse*. Tahap *Revise* dan *Retain* tidak diimplementasikan karena penelitian ini tidak melibatkan validasi pakar maupun pembaruan basis kasus secara otomatis.
4. Penelitian ini mengadaptasi kerangka kerja CRISP-DM sampai dengan tahap *evaluation*. Tahap *deployment* tidak dilakukan karena penelitian berfokus pada perancangan dan evaluasi performa model dalam lingkungan eksperimen.

## 1.5 Manfaat

Penelitian tugas akhir ini diharapkan dapat memberikan beberapa manfaat, antara lain sebagai berikut:

### 1.5.1 Manfaat Teoritis

Penelitian diharapkan dapat memberikan manfaat secara teoritis, diantaranya sebagai berikut:

1. Memberikan kontribusi dalam pengembangan pendekatan *Structural Semantic Retrieval* melalui integrasi representasi semantik berbasis transformer (Indo-LegalBERT) dan representasi struktural berbasis *Knowledge Graph* sebagai pendekatan alternatif dalam menangani keterbatasan model semantik konvensional dalam menangkap relasi kausal pada teks naratif panjang.
2. Menjadi referensi akademik dalam pengembangan sistem temu kembali informasi berbasis *Case-Based Reasoning* yang mengombinasikan kemiripan semantik dan struktural, khususnya pada dokumen dengan kompleksitas kontekss dan relasi yang tinggi.

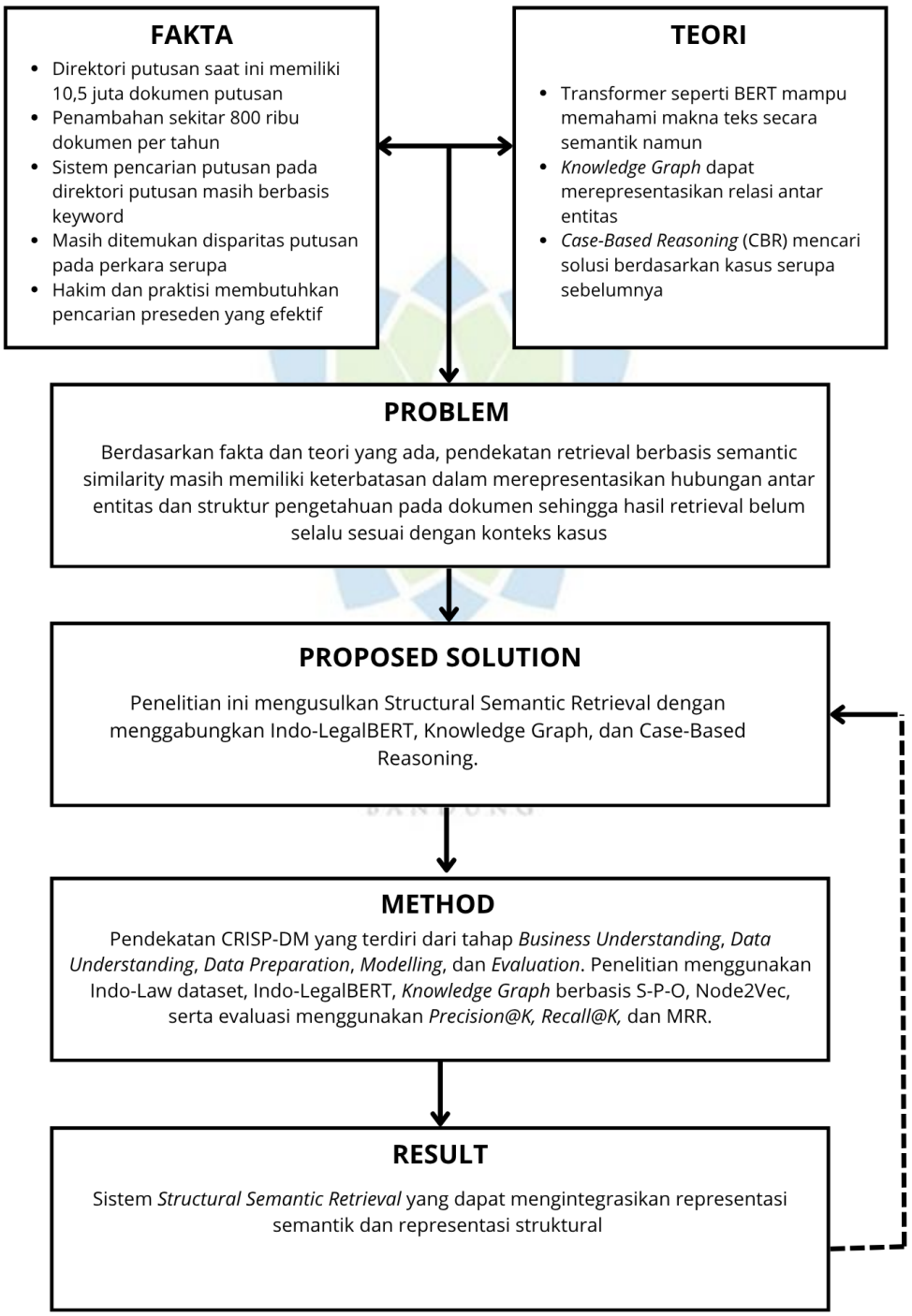
### 1.5.2 Manfaat Praktis

Penelitian ini juga diharapkan dapat memberikan manfaat praktis, diantaranya sebagai berikut:

1. Menghasilkan prototipe sistem *legal retrieval* yang dapat digunakan untuk menemukan dokumen putusan berdasarkan kemiripan semantik dan hubungan antar entitas hukum sebagai alternatif dari pendekatan pencarian berbasis kecocokan kata kunci.
2. Menjadi referensi dalam pengembangan sistem pencarian dokumen hukum yang dapat membantu proses penelusuran putusan dan argumentasi hukum bagi hakim, mahasiswa hukum, peneliti, maupun praktisi hukum.

## 1.6 Kerangka Pemikiran

Kerangka pemikiran pada penelitian ini berdasarkan fakta, literatur, masalah, solusi, metodologi yang digunakan, hingga hasil yang diharapkan dalam penelitian ini.




Gambar 1. 1 Kerangka Pemikiran

Gambar 1.1 menunjukkan alur pemikiran penelitian yang dimulai dari aspek fakta dan landasan teori, kemudian dirumuskan menjadi permasalahan, solusi yang diusulkan, metode, hingga hasil yang diharapkan. Pada bagian *fakta*, ditunjukkan bahwa dokumen putusan hukum memiliki struktur naratif yang kompleks dengan relasi antar entitas, serta banyak kasus yang memiliki kemiripan kata namun berbeda relasi sehingga menghasilkan makna hukum yang berbeda. Selain itu, meskipun sistem *Information Retrieval* telah berkembang dari *keyword matching* ke *semantic matching*, pendekatan tersebut masih memiliki keterbatasan dalam memahami struktur.

Pada bagian *teori*, dijelaskan bahwa model berbasis transformer cenderung mengalami *structural blindness* (buta struktur), yaitu tidak mampu menangkap relasi kausal antar entitas secara eksplisit. Sebaliknya, *Knowledge Graph* mampu merepresentasikan hubungan antar entitas dalam bentuk struktur graf (S–P–O), dan integrasi struktur graf terbukti dapat meningkatkan kualitas ekstraksi informasi. Berdasarkan fakta dan teori tersebut, dirumuskan *problem* bahwa sistem pencarian yang umum digunakan saat ini belum mampu menangkap relasi antar entitas secara eksplisit, sehingga pendekatan semantik saja sering gagal membedakan dokumen dengan alur yang berlawanan meskipun memiliki kosakata yang mirip.

Sebagai *proposed solution*, penelitian ini mengusulkan pendekatan *hybrid* yang menggabungkan *embedding* semantik dan representasi struktural berbasis *Knowledge Graph* dalam kerangka *Case-Based Reasoning* (CBR) untuk proses *retrieval* dokumen. Selanjutnya, pada bagian *method*, penelitian ini menerapkan kerangka kerja CRISP-DM dengan tahapan konstruksi *Knowledge Graph*, pembentukan *embedding* semantik dan struktural, serta proses pencarian berbasis kemiripan yang dievaluasi menggunakan metrik *Precision*, *Recall*, dan *MRR*. Hasil akhir yang diharapkan sistem mampu merekomendasikan dokumen yang relevan dengan peningkatan kinerja berdasarkan metrik evaluasi tersebut.

## 1.7 Sistematika Penulisan

Sistematika penulisan laporan memuat sistematika penulisan laporan tugas akhir dengan memberikan gambaran kandungan setiap bab, urutan penulisan, serta keterkaitan antara satu bab dengan bab lainnya dalam sebuah laporan tugas akhir. Berikut sistematika penulisan laporan tugas akhir.

**BAB I PENDAHULUAN**

Bab ini terdiri dari latar belakang, rumusan masalah, tujuan penelitian, batasan masalah, kerangka pemikiran, dan sistematika penulisan.

**BAB II TINJAUAN PUSTAKA**

Bab ini membahas literatur atau penelitian terdahulu, konsep dan teori, model yang digunakan, dan rumus terkait yang menjadi landasan dalam proses analisis permasalahan dan kebutuhan penelitian.

**BAB III METODOLOGI PENELITIAN**

Bab ini berisi penjelasan langkah-langkah dan teknik yang diterapkan dalam penelitian, diuraikan secara sistematis dan terstruktur.

**BAB IV HASIL DAN PEMBAHASAN**

Bab ini memaparkan dua hal utama, yaitu temuan atau hasil penelitian berdasarkan langkah-langkah penelitian yang telah dilakukan, dan pembahasan hasil atau temuan penelitian sebagai jawaban terhadap rumusan masalah penelitian.

**BAB V KESIMPULAN DAN SARAN**

Bab ini berfokus pada penarikan simpulan dari hasil penelitian yang diperoleh serta menjawab pertanyaan penelitian, serta memberikan saran yang dapat dilakukan penelitian selanjutnya untuk meningkatkan kualitas penelitian.