BAB I
PENDAHULUAN

1.1 Latar Belakang

Dalam praktik penegakan hukum pidana, ketepatan dalam menentukan pasal yang sesuai dengan fakta perkara merupakan aspek yang sangat penting karena menjadi dasar dalam proses penuntutan dan pemidanaan. Kesalahan dalam penerapan pasal dapat menimbulkan ketidakpastian hukum, memengaruhi keadilan putusan, serta berujung pada upaya hukum lanjutan hingga tingkat kasasi. Data Mahkamah Agung Republik Indonesia tahun 2024 menunjukkan bahwa dari 20.317 perkara kasasi yang diputus, terdapat 2.111 perkara yang dikabulkan. Sebagian perkara yang dikabulkan tersebut berkaitan dengan kesalahan penerapan hukum atau kesalahan substansi hukum dalam putusan sebelumnya [1]. Beberapa penelitian juga menunjukkan adanya kasus di mana pasal yang diterapkan tidak sepenuhnya sesuai dengan fakta perkara sehingga Mahkamah Agung melakukan koreksi terhadap putusan tersebut [2], [3]. Kasus-kasus tersebut merupakan contoh nyata bagaimana perbedaan redaksi dan unsur delik dalam Kitab Undang-Undang Hukum Pidana (KUHP) dapat menimbulkan kesalahan interpretasi hukum. Saat ini KUHP masih menjadi rujukan utama dalam penanganan perkara pidana di Indonesia dan memiliki struktur normatif yang kompleks, dengan banyak pasal yang saling beririsan secara redaksional maupun konseptual [4].

Perkembangan pemrosesan bahasa alami telah mengubah cara mesin memahami teks melalui pemanfaatan representasi semantik. Representasi semantik menggambarkan makna dan hubungan konseptual dalam teks sehingga sistem dapat memahami keterkaitan makna yang tidak dinyatakan secara eksplisit. Dalam Natural Language Processing (NLP), representasi semantik dibangun menggunakan embedding yang merepresentasikan kata atau kalimat ke dalam bentuk vektor numerik pada ruang multidimensi [5]. Pada vektor tersebut, unsur teks dengan makna serupa cenderung berada pada posisi yang berdekatan, sehingga analisis semantik dapat dilakukan dengan membandingkan jarak atau kemiripan antar vektor untuk mengukur kesamaan makna antar kalimat atau dokumen [6]. Efektivitas metode ini dibuktikan melalui penelitian bilingual

Indonesia–Inggris yang memanfaatkan *Universal Sentence Encoder (USE)* dan *Facebook AI Similarity Search (FAISS)*. Studi tersebut mencapai *F1-score* sebesar 96%, menunjukkan bahwa model mampu mengenali kemiripan makna antar teks secara presisi bahkan pada dua bahasa yang berbeda [7] .

Berdasarkan pemanfaatan representasi semantik untuk mengukur kesamaan makna antar teks, kajian *text matching* dalam *Natural language processing* (NLP) mengklasifikasikan pendekatan pencocokan makna ke dalam beberapa teori utama. Secara umum, *lexical-based matching* dan *semantic-based* matching diposisikan sebagai komponen fundamental karena menjadi dasar dalam menilai kesamaan teks. Pendekatan leksikal berfokus pada kesesuaian istilah dan frekuensi kata, sedangkan pendekatan semantik memanfaatkan representasi vektor kontekstual untuk menangkap kesamaan makna yang tidak selalu ditandai oleh kesamaan kosakata [8]. Selain dua komponen utama tersebut, literatur juga mengidentifikasi pendekatan pendukung seperti *Knowledge-based matching* yang memanfaatkan struktur pengetahuan seperti ontologi untuk merepresentasikan hubungan konseptual secara terstruktur [8], [9], pencocokan sintaktis yang menekankan struktur gramatikal kalimat, serta *context-aware matching* yang mempertimbangkan konteks penggunaan dan karakteristik domain tertentu [8].

Dalam konteks Bahasa Indonesia, pemahaman semantik dibangun melalui relasi makna seperti sinonim (persamaan makna), hiponim (hubungan makna umum dan khusus) dan hubungan antar konsep dalam struktur bahasa, sebagaimana ditunjukkan oleh penelitian Yessy (2023) [10]. Pengembangan pemodelan bahasa Indonesia berkembang pesat melalui pengembangan model *embedding* seperti IndoBERT dan IndoSBERT. Kedua model ini dirancang untuk menangkap ciri khas bahasa Indonesia secara lebih akurat dibandingkan model generik yang tidak dilatih secara khusus pada bahasa Indonesia [11], [12]. Penelitian Abdul Rozaq et al. (2025) menunjukkan bahwa IndoSBERT dapat meningkatkan kinerja sistem analisis teks hukum dengan akurasi rekomendasi hukum sebesar 76,66% ketika diuji menggunakan lebih dari dua ribu dokumen hukum [13].

Kinerja IndoBERT terbukti meningkat signifikan ketika dilakukan pelatihan awal dan penyesuaian model pada domain hukum. Penelitian yang

menggabungkan arsitektur IndoBERT dengan lapisan *Conditional Random Field* untuk memahami struktur pasal, ayat, dan rujukan hukum menghasilkan nilai F1 sebesar 92,3% [1]. Studi lain mengenai efektivitas model Legal-BERT menunjukkan bahwa model ini mencapai nilai *Macro F1* sebesar 69,5 persen, yang merupakan nilai tertinggi dibandingkan model pembanding lainnya seperti IndoBERT dalam pengolahan teks hukum [14]. Penelitian mengenai klasifikasi dokumen putusan perdata juga memperjelas keunggulan model yang dilatih secara khusus pada dokumen hukum pidana Indonesia [15]. IndoLegalBERT, yang memanfaatkan arsitektur *transformer* dan dilatih menggunakan ribuan dokumen hukum Indonesia, terbukti menghasilkan performa lebih baik dibandingkan Legal-BERT [16]. Hal ini disebabkan oleh kemampuan IndoLegalBERT dalam memahami terminologi hukum, pola redaksi regulatif, serta struktur penalaran normatif dalam peraturan perundang-undangan Indonesia.

Model IndoLegalBERT terbukti unggul dalam menangkap makna kontekstual, namun pendekatan semantik murni memiliki keterbatasan ketika dihadapkan pada dokumen hukum yang panjang dan memiliki redaksi yang sangat mirip. Penelitian di bidang *information Retrieval* menunjukkan bahwa pendekatan *Hybrid Retrieval* yang menggabungkan metode leksikal seperti BM25 dengan *dense Retrieval* berbasis model *Transformer* mampu menghasilkan kinerja yang lebih optimal. Penelitian internasional menggunakan Pyserini melaporkan peningkatan konsisten pada metrik nDCG@10 (relevansi) dan MAP (ketepatan) melalui kombinasi BM25 dan *dense Retrieval* [17]. Pendekatan serupa juga diperkuat oleh temuan empiris yang menunjukkan bahwa *Hybrid dense–sparse Retrieval* mampu mencapai peningkatan recall (proporsi) sekitar 25 – 30 % dan peningkatan MAP (ketepatan) sekitar 12% dibandingkan pendekatan berbasis leksikal saja [18]. Selain itu, penelitian pada *domain-specific question answering* menunjukkan bahwa integrasi BM25 dengan *retriever* semantik yang telah di-*fine-tune* menghasilkan skor nDCG (relevansi) yang lebih tinggi (0,845) dibandingkan metode semantik saja (0,828) maupun BM25 saja (0,640), sehingga relevan untuk penelusuran teks hukum [19]

Berdasarkan temuan penelitian sebelumnya, pendekatan *Hybrid Retrieval* terbukti mampu meningkatkan kualitas penelusuran informasi dibandingkan

penggunaan metode leksikal atau semantik secara terpisah. Oleh karena itu, penelitian ini mengusulkan kombinasi BM25 dan IndoLegalBERT untuk rekomendasi pasal KUHP berdasarkan fakta perkara. BM25 digunakan untuk menangkap kecocokan terminologi hukum, sedangkan IndoLegalBERT digunakan untuk memahami kesamaan makna. Melalui kombinasi tersebut, sistem diharapkan mampu menghasilkan rekomendasi pasal yang lebih relevan dan akurat.

1.2 Rumusan Masalah

Rumusan masalah bertujuan untuk memfokuskan penelitian pada permasalahan utama agar tujuan penelitian dapat tercapai.

1. Bagaimana mengimplementasikan pendekatan *Hybrid Retrieval* (IndoLegalBERT dan BM25) yang optimal untuk memetakan fakta kasus kedalam label pasal KUHP yang relevan?
2. Bagaimana kinerja model IndoLegalBERT dalam merekomendasikan pasal KUHP menggunakan metrik evaluasi *Retrieval*?

1.3 Tujuan

Tujuan penelitian ini untuk memberikan arah yang jelas terhadap pelaksanaan penelitian agar hasil yang diperoleh sesuai dengan fokus kajian.

1. Mengimplementasikan pendekatan *Hybrid Retrieval* yang mengintegrasikan IndoLegalBERT dan algoritma BM25 untuk memetakan fakta kasus ke dalam pasal KUHP yang relevan, dengan menggabungkan pencocokan leksikal dan pemahaman semantik guna mengoptimalkan sistem rekomendasi pasal hukum.
2. Mengukur kinerja model IndoLegalBERT menggunakan metrik evaluasi Retrieval, yaitu *Top-K Accuracy* (Top-1 dan Top-3) untuk melihat ketepatan hasil teratas yang berhasil ditemukan model, *Mean Average Precision* (MAP) untuk mengukur rata-rata presisi hasil retrieval, serta *Mean Reciprocal Rank* (MRR) untuk mengevaluasi posisi peringkat jawaban relevan pertama yang diberikan model.

1.4 Batasan Masalah

Batasan masalah digunakan untuk membatasi ruang lingkup penelitian agar pembahasan tetap fokus dan sesuai dengan tujuan penelitian.

1. Data yang digunakan berupa teks pasal dari Kitab Undang-Undang Hukum Pidana (KUHP) yang berisi 569 pasal, serta teks fakta kasus dalam bentuk narasi pendek atau ringkasan kasus.
2. Model bahasa yang digunakan adalah IndoLegalBERT, tanpa dilakukan *fine-tuning* tambahan terhadap model hukum lainnya, dimana evaluasi model dibatasi pada pengukuran nilai *cosine similarity* untuk menentukan tingkat kemiripan makna, serta penggunaan metrik evaluasi *Retrieval*.
3. Penelitian ini hanya menggunakan satu pasal utama sebagai *ground truth* untuk setiap fakta kasus berdasarkan surat putusan Mahkamah Agung. Pasal yang bersifat junto, pemberatan, peringanan, atau pasal pendukung lainnya tidak termasuk dalam proses evaluasi.
4. Sistem rekomendasi hanya menampilkan daftar tiga pasal dengan skor kemiripan tertinggi tanpa mempertimbangkan aspek yurisprudensi atau tafsir hukum lanjutan.
5. Penelitian difokuskan pada tindak pidana yang diatur dalam KUHP sebagai ruang lingkup utama penelitian, sehingga pembahasan diarahkan pada pasal-pasal pidana umum dalam KUHP.
6. Metode penelitian yang digunakan adalah Cross Industry Standard Process for Data Mining (CRISP-DM), dengan pelaksanaan tahapan dibatasi hingga fase evaluasi model. Tahap deployment tidak termasuk dalam ruang lingkup penelitian ini.

1.5 Manfaat

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

1.5.1 Manfaat Akademik

Penelitian ini akan memberikan manfaat bagi akademik, diantaranya:

1. Memberikan kontribusi ilmiah dalam pengembangan ilmu komputer, khususnya pada bidang *Natural language processing* (NLP) untuk domain hukum Indonesia melalui penerapan model IndoLegalBERT.
2. Memperkaya literatur mengenai efektivitas metode *Hybrid Retrieval* (penggabungan *Semantic Similarity* dan algoritma leksikal BM25) dalam mengatasi tantangan kompleksitas bahasa pada dokumen hukum nasional.

1.5.2 Manfaat bagi Praktisi Hukum

Penelitian ini akan memberikan manfaat bagi praktisi hukum, diantaranya:

1. Mendukung terciptanya proses penegakan hukum yang lebih transparan dan objektif karena hasil analisis didasarkan pada data dan kedekatan makna sistematis, bukan semata interpretasi subjektif individu.
2. Membantu praktisi hukum, khususnya jaksa, dalam mempercepat proses identifikasi pasal pada perkara pidana melalui rujukan awal yang relevan dengan teks fakta kasus, sehingga dapat meminimalisir kesalahan penerapan pasal.

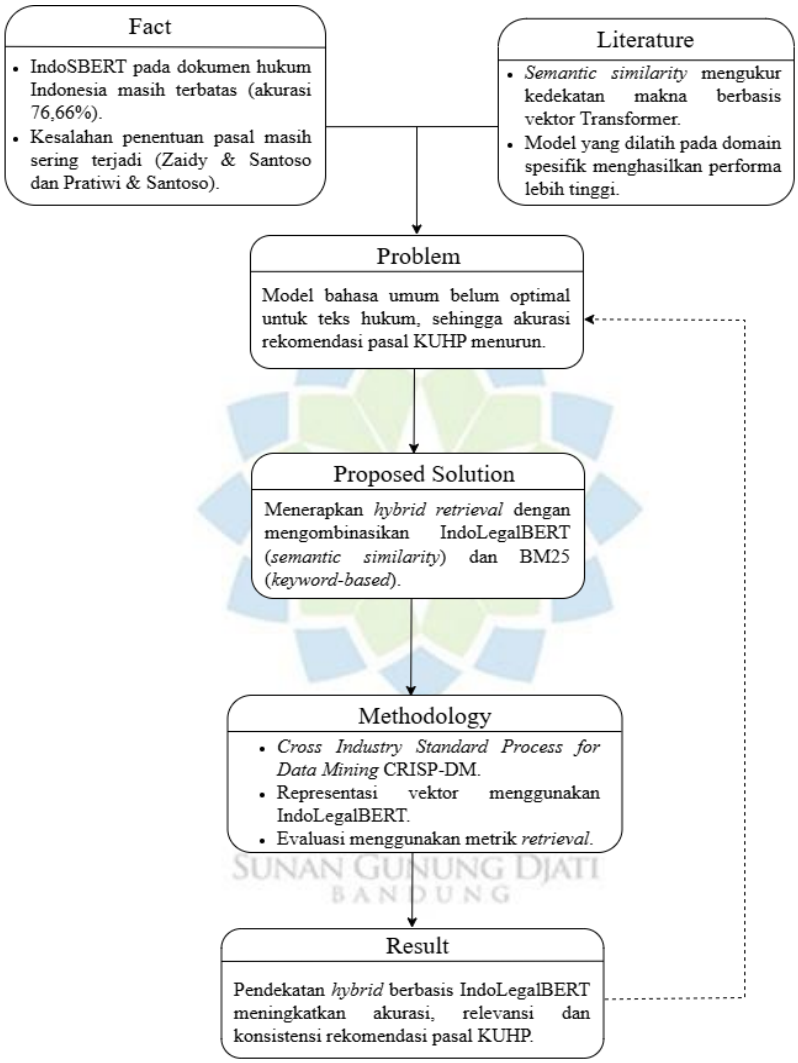
1.5.3 Manfaat bagi Peneliti

Penelitian ini akan memberikan manfaat bagi peneliti, diantaranya:

1. Menjadi referensi empiris bagi peneliti selanjutnya yang ingin mengembangkan sistem rekomendasi hukum atau sistem pencarian informasi (*Information Retrieval*) berbasis model *Transformer*.
2. Memberikan wawasan teknis mengenai implementasi pendekatan *Hybrid* dalam menyeimbangkan presisi pencarian kata kunci dan pemahaman konteks makna pada data hukum yang tidak terstruktur.

1.6 Kerangka Pemikiran

Kerangka pemikiran ini menjelaskan alur penelitian yang disusun berdasarkan fakta, kajian literatur, perumusan masalah, solusi yang ditawarkan, serta metodologi penelitian yang diterapkan.



Gambar 1. 1 Kerangka Pemikiran

Gambar 1.1 menggambarkan kerangka berpikir penelitian yang disusun secara sistematis mulai dari fakta empiris, landasan literatur, hingga hasil yang diharapkan. Pada tahap fakta, ditunjukkan bahwa model bahasa umum seperti IndoSBERT masih menunjukkan keterbatasan ketika diterapkan pada dokumen hukum Indonesia, yang tercermin dari capaian akurasi yang relatif rendah serta masih ditemukannya kesalahan pemilihan pasal dalam praktik peradilan pidana.

Fakta ini diperkuat oleh temuan empiris dalam putusan pengadilan yang menunjukkan ketidaksesuaian antara fakta perkara dan pasal yang diterapkan. Fakta tersebut kemudian dikaitkan dengan kajian literatur yang menegaskan bahwa pendekatan *Semantic Similarity* berbasis *Transformer* bekerja dengan merepresentasikan teks ke dalam vektor semantik, serta menunjukkan bahwa performa model cenderung meningkat signifikan apabila dilatih pada korpus yang spesifik terhadap domain tertentu, seperti domain hukum.

Berdasarkan keterkaitan antara fakta dan literatur tersebut, masalah penelitian dirumuskan pada ketidakoptimalan model bahasa umum dalam merepresentasikan karakteristik teks hukum, yang berdampak pada menurunnya akurasi rekomendasi pasal KUHP. Sebagai solusi, penelitian ini mengusulkan penerapan pendekatan *Hybrid Retrieval* dengan mengombinasikan IndoLegalBERT sebagai representasi semantik dan BM25 sebagai pencocokan berbasis kata. Fakta perkara dan pasal KUHP direpresentasikan dalam bentuk vektor dan dievaluasi menggunakan metrik akurasi *Retrieval*. Hasil penelitian menunjukkan bahwa pendekatan ini mampu meningkatkan akurasi dan konsistensi rekomendasi pasal, sehingga lebih relevan dan efektif dalam mendukung analisis perkara pidana.

1.7 Sistematika Penulisan

Sistematika penulisan laporan menjelaskan susunan penulisan tugas akhir yang memuat gambaran isi setiap bab, urutan penyajian, serta keterkaitan antara satu bab dengan bab lainnya dalam sebuah laporan tugas akhir.

BAB I PENDAHULUAN

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

BAB II KAJIAN LITERATUR

Bab ini membahas penelitian terdahulu, konsep-konsep dasar, teori, model, serta rumus yang digunakan sebagai landasan teoretis dalam menganalisis permasalahan sesuai dengan topik penelitian.

BAB III METODOLOGI PENELITIAN

Bab ini menjelaskan metode penelitian yang digunakan, meliputi tahapan penelitian dan teknik yang diterapkan, yang diuraikan secara sistematis dan terstruktur.

BAB IV HASIL DAN PEMBAHASAN

Bab ini memuat pemaparan hasil penelitian yang diperoleh berdasarkan tahapan metodologi yang telah dilaksanakan, termasuk hasil pengujian dan evaluasi sistem. Selanjutnya, dilakukan pembahasan terhadap hasil tersebut dengan mengaitkannya pada tujuan penelitian, rumusan masalah, serta kajian literatur untuk menilai kinerja dan efektivitas metode yang diusulkan.

BAB V SIMPULAN DAN SARAN

Bab ini menyajikan simpulan yang dirumuskan berdasarkan hasil analisis dan pembahasan penelitian sebagai jawaban atas rumusan masalah. Selain itu, bab ini juga memuat saran yang ditujukan untuk pengembangan sistem maupun penelitian selanjutnya.