BAB I
PENDAHULUAN

1.1 Latar Belakang

Negara Kesatuan Republik Indonesia sebagai negara hukum menjunjung tinggi asas fiksi hukum atau *ignorantia juris non excusat*. Asas ini menetapkan aturan mutlak bahwa setiap warga negara dianggap sudah mengetahui dan memahami suatu undang-undang sejak peraturan tersebut disahkan. Akibatnya, ketidaktahuan terhadap hukum tidak bisa dijadikan alasan pemaaf dalam proses peradilan. Namun, pada kenyataannya, terdapat kesenjangan yang sangat besar antara aturan ideal tersebut dengan kemampuan literasi masyarakat.

Kesenjangan ini terbukti secara faktual melalui berbagai fenomena sosial dan data statistik. Sebagai contoh, dalam polemik Undang-Undang Cipta Kerja, survei Saiful Mujani Research and Consulting (SMRC) menunjukkan bahwa 74% masyarakat belum mengetahui adanya Rancangan Undang-Undang Cipta Kerja [1]. Temuan tersebut menunjukkan masih rendahnya pengetahuan publik terhadap regulasi yang pada saat itu tengah menjadi perhatian nasional. Sejalan dengan temuan tersebut, survei Litbang Kompas pada tahun 2023 mencatat bahwa 60,5% publik menilai Undang-Undang Cipta Kerja tidak mewakili aspirasi mereka [2]. Kondisi tersebut menunjukkan adanya kesenjangan antara pembentukan regulasi dengan pemahaman dan persepsi masyarakat terhadap substansi peraturan. Kompleksitas struktur dan bahasa dalam dokumen perundang-undangan dapat menjadi salah satu faktor yang menyulitkan masyarakat awam dalam memahami substansi aturan secara utuh. Ketidakpahaman terhadap isi peraturan tersebut berpotensi memengaruhi kemampuan masyarakat dalam memperoleh informasi hukum yang benar serta membentuk pemahaman yang tepat mengenai hak dan kewajiban yang diatur dalam peraturan perundang-undangan.

Berdasarkan prinsip hukum yang berlaku, ketidakmampuan masyarakat dalam mencerna teks hukum yang rumit dapat menghambat terwujudnya partisipasi publik yang bermakna. Hal ini terbukti secara faktual ketika Mahkamah Konstitusi, melalui Putusan Nomor 91/PUU-XVIII/2020, menyatakan undang-undang tersebut inkonstitusional bersyarat. Dalam amar putusannya, Mahkamah Konstitusi secara

eksplisit menyatakan bahwa Undang-Undang Nomor 11 Tahun 2020 tentang Cipta Kerja bertentangan dengan Undang-Undang Dasar Negara Republik Indonesia Tahun 1945 dan tidak mempunyai kekuatan hukum mengikat secara bersyarat sepanjang tidak dimaknai dilakukan perbaikan dalam jangka waktu dua tahun sejak putusan tersebut diucapkan. Rangkaian fenomena ini mempertegas bahwa rendahnya aksesibilitas bahasa hukum berisiko mendisrupsi stabilitas nasional, menurunkan legitimasi produk legislasi, serta memperlebar jurang pemenuhan hak fundamental masyarakat terhadap keadilan [3].

Berdasarkan Laporan Situasi Hak-Hak Digital Indonesia oleh SAFEnet, ketidakjelasan linguistik pada pasal multitafsir UU ITE terus memakan korban. Sepanjang 2023 tercatat 89 kasus pemidanaan terhadap warganet sipil yang mayoritas dipicu oleh perbedaan interpretasi frasa "pencemaran nama baik" pada Pasal 27 [4]. Di sisi lain, Laporan Status Literasi Digital Indonesia oleh Kementerian Kominfo menunjukkan bahwa pilar Keamanan Digital masih menjadi yang terendah dengan skor 3,12 dari skala 5,00. Rendahnya literasi ini berkorelasi dengan temuan bahwa lebih dari 60% masyarakat belum mampu memahami konsekuensi hukum dari kebijakan privasi sehingga gagal mempraktikkan hak perlindungan data berdasarkan UU PDP [5]. Kondisi tersebut semakin diperkuat dengan hadirnya KUHP Baru yang memuat 624 pasal pemidanaan, sehingga pemerintah menetapkan masa transisi selama tiga tahun untuk sosialisasi guna mencegah kriminalisasi akibat ketidaktahuan. Data empiris tersebut menunjukkan bahwa kompleksitas bahasa hukum telah memicu kerentanan hukum dan sipil di berbagai sektor.

Fenomena ini sejalan dengan data tingkat literasi nasional laporan PISA (OECD) secara konsisten menempatkan kemampuan membaca siswa Indonesia di peringkat bawah, di mana sekitar 70% siswa belum mencapai kompetensi minimum untuk memahami teks panjang [6]. Sementara itu, dokumen hukum memiliki tingkat keterbacaan yang sangat rendah dan secara akademis mensyaratkan tingkat pendidikan tinggi untuk memahaminya. Mengingat Angka Partisipasi Kasar (APK) Perguruan Tinggi di Indonesia baru mencapai kisaran 31% (BPS) [7], maka secara statistik mayoritas masyarakat Indonesia menghadapi hambatan signifikan dalam

pemenuhan hak fundamental mereka, yakni Akses terhadap Keadilan (*Access to Justice*), akibat ketidakmampuan memahami dokumen hukum secara mandiri [8].

Hambatan utama tersebut bersumber dari karakteristik dokumen perundang-undangan di Indonesia yang secara tradisional disusun menggunakan ragam bahasa hukum (*legalese*) yang hermetis. Karakteristik ini tercermin dalam penggunaan struktur kalimat majemuk bertingkat, terminologi teknis yang kaku, serta sistem referensi silang antar pasal yang rumit bagi pembaca non-ahli. Kompleksitas linguistik ini telah diidentifikasi sebagai barikade utama dalam upaya peningkatan transparansi hukum, sebagaimana juga ditemukan dalam berbagai studi linguistik forensik dan *Natural Legal Language Processing* [9].

Dalam konteks perkembangan teknologi, tantangan keterbacaan ini dapat ditangani melalui disiplin *Natural Language Processing* (*NLP*), khususnya tugas *Automatic Text Simplification* (*ATS*). Berbeda dengan peringkasan (*text summarization*) yang mereduksi informasi, *ATS* bertujuan melakukan transformasi ulang teks (*rewriting*) untuk menurunkan kompleksitas leksikal dan sintaksis tanpa mengubah makna inti dari teks hukum. Studi terkini menunjukkan bahwa pendekatan berbasis *Deep Learning* semakin dominan dalam pengembangan sistem penyederhanaan hukum karena kemampuannya menangkap konteks semantik jangka panjang yang tidak dapat dilakukan oleh metode statistik konvensional [10]. Keberhasilan arsitektur *Transformer* (seperti T5 dan GPT) dalam memetakan kompleksitas dokumen legal menuju bahasa awam juga telah dibuktikan secara global, di mana model generatif ini secara konsisten mencatatkan peningkatan luar biasa pada metrik indeks keterbacaan (*readability indices*) pada berbagai penelitian penyederhanaan hukum lintas bahasa [11].

Meskipun demikian, penerapan *ATS* pada dokumen hukum Bahasa Indonesia masih menghadapi tantangan fundamental, terutama terkait kelangkaan sumber daya data (*data scarcity*) berupa korpus paralel dan isu validitas semantik. Kesalahan kecil dalam penyederhanaan hukum dapat berimplikasi serius terhadap makna normatif. Mengingat evaluasi preservasi makna pada domain legal memiliki tingkat sensitivitas dan risiko cacat logika yang jauh lebih krusial dibandingkan dengan domain teks umum [10], maka evaluasi model tidak cukup hanya

mengandalkan metrik keterbacaan standar. Penilaian harus dilakukan secara komprehensif dengan melibatkan validasi berbasis pakar hukum (*subject matter experts*) guna menjamin keamanan substansi [12] dan pakar linguistik guna menilai keterbacaan serta ketepatan bahasa..

Untuk menjawab tantangan tersebut, penelitian ini mengusulkan pendekatan komprehensif dengan mengimplementasikan model IndoT5 (Indonesian Text-to-Text Transfer Transformer). Pemilihan IndoT5 didasarkan pada dua alasan strategis secara garis besar, yaitu keunggulan arsitektur Sequence-to-Sequence (Encoder-Decoder) yang sangat efektif untuk tugas pembangkitan bahasa (Natural Language Generation) [13], serta keandalannya sebagai model monolingual yang dilatih spesifik pada korpus bahasa Indonesia sehingga lebih akurat dalam menangkap morfologi lokal dibandingkan model *multilingual* [14]. Secara empiris, keluarga algoritma IndoT5 telah terbukti memiliki kapabilitas yang superior dalam menghasilkan teks parafrase berbahasa Indonesia dengan metrik pelestarian semantik yang sangat solid dan gaya tata bahasa yang natural [15]. Penelitian ini akan berfokus pada pembangunan dataset paralel hukum berskala menengah (1623 pasang kalimat) dan penerapan evaluasi hibrida (metrik otomatis dan validasi pakar) untuk menghasilkan sistem penyederhanaan yang akurat, aman secara normatif, dan mudah dipahami oleh masyarakat luas.

1.2 Rumusan Masalah

Berdasarkan latar belakang di atas, rumusan masalah dalam penelitian ini adalah:

1. Bagaimana mengimplementasikan dan melakukan *fine-tuning* pada model IndoT5 untuk tugas penyederhanaan teks hukum (*legal text simplification*) bahasa Indonesia?
2. Bagaimana kinerja model IndoT5 dalam menghasilkan teks sederhana berdasarkan metrik evaluasi otomatis (*BLEU* dan *SARI*) dibandingkan dengan teks aslinya?
3. Bagaimana validasi pakar hukum dan pakar linguistik terhadap hasil penyederhanaan model IndoT5?

1.3 Tujuan Penelitian

Tujuan yang ingin dicapai dalam penelitian ini adalah:

1. Membangun dan menerapkan model IndoT5 yang telah melalui proses *fine-tuning* menggunakan dataset paralel teks hukum untuk menghasilkan penyederhanaan teks yang efektif.
2. Mengukur dan mengevaluasi kinerja model IndoT5 menggunakan metrik evaluasi standar (*BLEU* dan *SARI*) untuk mengetahui kualitas prediksi model.
3. Menganalisis hasil validasi pakar terhadap hasil penyederhanaan model IndoT5.

1.4 Manfaat Penelitian

Adapun hasil dari penelitian ini diharapkan dapat memberikan manfaat yang signifikan, seperti:

a. Manfaat Akademis
   1) Memberikan kontribusi keilmuan di bidang *Natural Language Processing* (*NLP*), khususnya dalam penerapan *Pre-trained Language Models* (*PLM*) untuk Bahasa Indonesia.
   2) Menjadi referensi bagi penelitian selanjutnya mengenai *Automatic Text Simplification* (*ATS*) pada domain spesifik (hukum) yang masih jarang dieksplorasi di Indonesia.

b. Manfaat Bagi Penulis
   1) Mendapatkan pemahaman mendalam mengenai arsitektur Transformer dan mekanisme model IndoT5.
   2) Meningkatkan kemampuan teknis dalam pengolahan data teks, *fine-tuning* model *Machine Learning*, dan evaluasi sistem *NLP*.

c. Manfaat Bagi Masyarakat
   1) Membantu masyarakat awam untuk lebih mudah memahami isi dokumen hukum atau undang-undang yang seringkali menggunakan bahasa yang kompleks dan kaku.
   2) Meningkatkan literasi hukum di masyarakat dengan menyediakan akses terhadap informasi hukum yang lebih sederhana dan mudah dicerna.

1.5 Batasan Masalah

1.5.1 Ruang Lingkup Masalah

Agar penelitian tetap fokus dan dapat diselesaikan dalam kerangka waktu Tugas Akhir, penulis menetapkan batasan masalah sebagai berikut:

1. **Sumber Data**: Data teks yang digunakan bersumber dari dokumen undang-undang Indonesia, yang difokuskan pada bagian "Batang Tubuh" (pasal-pasal). Sumber spesifik meliputi UU No. 1 Tahun 2023 (KUHP Baru), UU No. 1 Tahun 2024 (Perubahan Kedua UU ITE), UU No. 6 Tahun 2023 (Cipta Kerja), dan UU Perlindungan Data Pribadi (UU PDP). Bagian "Penjelasan" dan lampiran undang-undang tidak disertakan dalam dataset.
2. **Metode Pembangunan Dataset**: Proses akuisisi data diawali dengan pengumpulan dan ekstraksi teks secara manual dari dokumen perundang-undangan asli. Tahapan anotasi untuk membangun korpus paralel dilakukan melalui metode *Semi-Automated Annotation*. Metode ini memanfaatkan *Large Language Model* (LLM) untuk menghasilkan kandidat penyederhanaan awal, yang selanjutnya diwajibkan melalui tahap verifikasi dan perbaikan manual oleh manusia (*human-in-the-loop*) untuk mencegah terjadinya distorsi atau halusinasi pada makna hukum.
3. **Volume Dataset**: Jumlah korpus paralel yang digunakan dalam proses pelatihan dan pengujian dibatasi secara eksklusif pada data yang telah dibersihkan sebanyak 1.623 pasangan kalimat. Dataset ini kemudian didistribusikan secara acak menggunakan rasio 80:10:10 menjadi 1.298 data latih, 162 data validasi, dan 163 data uji.
4. **Lingkup Hasil (Luaran)**: Hasil keluaran sistem adalah teks yang disederhanakan secara leksikal dan sintaksis (*Text Simplification*). Penelitian ini tidak mencakup tugas peringkasan dokumen (*Text Summarization*) ataupun penafsiran hukum (*Legal Interpretation*) yang bersifat mengikat secara yurisprudensi.
5. **Arsitektur Model**: Ruang lingkup pemodelan dibatasi pada arsitektur *Transformer* berjenis *Sequence-to-Sequence* (seq2seq) menggunakan *pre-trained model* IndoT5 varian Base (IndoT5-base). Proses *fine-tuning* dan

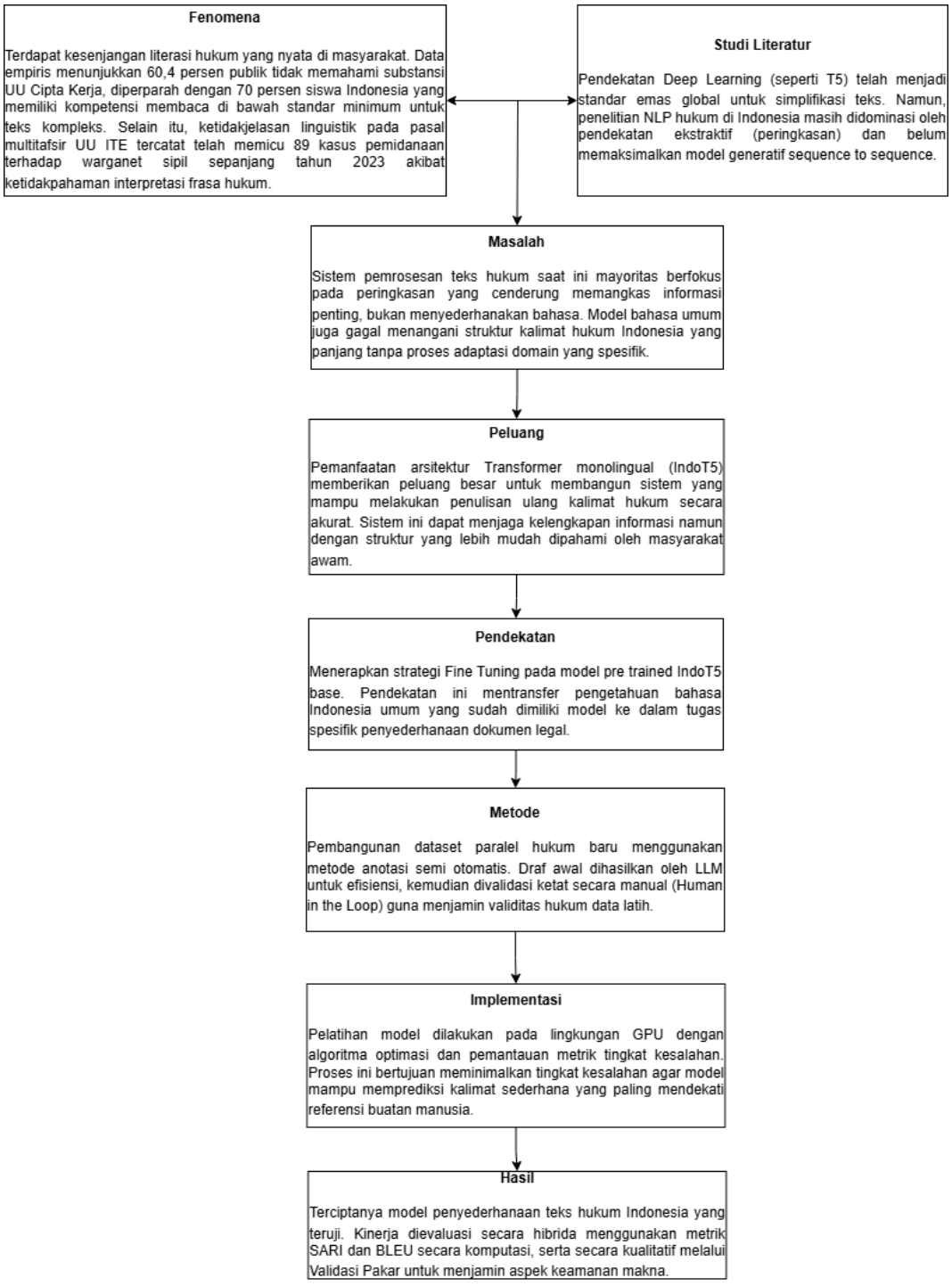
penyesuaian hiperparameter dikonfigurasi secara spesifik hanya untuk mereduksi kompleksitas sintaksis teks hukum. Penelitian ini tidak ditujukan untuk variasi tugas pemrosesan bahasa alami lainnya dan tidak membandingkan kinerja dengan varian arsitektur di luar keluarga T5 atau ukuran model lainnya (*small/large*).

6. **Metode Evaluasi**: Evaluasi manusia (*human evaluation*) oleh pakar tidak dilakukan pada seluruh data uji, melainkan menggunakan teknik *statistical sampling* untuk menentukan jumlah sampel representatif yang akan divalidasi guna mengukur metrik *Adequacy*, *Simplicity*, dan *Fluency*.

1.5.2 Keterbatasan Penelitian

Penelitian ini memiliki keterbatasan pada ukuran korpus paralel yang digunakan untuk proses pelatihan model. Data latih yang terdiri atas 1.298 pasang kalimat masih tergolong terbatas untuk mengoptimalkan kemampuan arsitektur IndoT5 secara menyeluruh. Keterbatasan tersebut disebabkan oleh minimnya ketersediaan dokumen hukum berbahasa Indonesia yang telah disederhanakan secara manual dan dapat digunakan sebagai korpus paralel berkualitas. Oleh karena itu, model yang dihasilkan dalam penelitian ini lebih tepat dipandang sebagai bukti awal (**proof of concept**) bahwa arsitektur IndoT5 memiliki potensi untuk diterapkan pada tugas penyederhanaan teks hukum. Dengan ketersediaan korpus paralel yang lebih besar dan beragam, performa model masih berpeluang untuk ditingkatkan, baik dalam mempertahankan makna hukum maupun menghasilkan teks yang lebih mudah dipahami oleh masyarakat.

1.6 Kerangka Berpikir



Gambar 1 1 Kerangka Berpikir

Kerangka berpikir penelitian ini diawali dengan mengkorelasikan fenomena nyata di masyarakat berupa rendahnya literasi hukum dan tingkat kemampuan membaca yang memicu tingginya kasus pemidanaan warganet dengan studi

literatur yang menunjukkan bahwa sistem pemrosesan teks di Indonesia masih didominasi oleh metode peringkasan ekstraktif. Korelasi antara fenomena dan literatur tersebut memunculkan rumusan masalah utama mengenai kegagalan sistem komputasi umum dalam menangani struktur kalimat hukum tanpa memangkas informasi penting di dalamnya. Situasi tersebut membuka peluang strategis untuk memanfaatkan arsitektur IndoT5 guna menyusun ulang kalimat kompleks secara generatif agar susunannya lebih mudah dipahami oleh masyarakat awam tanpa mereduksi makna absolut.

Pendekatan yang dipilih untuk merealisasikan peluang tersebut adalah teknik pembaruan bobot pada model pre trained IndoT5 base agar mesin beradaptasi dengan domain spesifik. Pendekatan tersebut dijalankan menggunakan metode pembangunan dataset paralel menggunakan anotasi semi otomatis yang divalidasi oleh manusia secara ketat. Data tersebut kemudian diimplementasikan ke dalam proses pelatihan komputasi tingkat tinggi untuk menekan tingkat kesalahan prediksi kata. Keseluruhan alur kerangka ini bermuara pada hasil pembentukan model akhir yang diukur tingkat keberhasilannya secara komprehensif menggunakan rumusan evaluasi matematis SARI dan BLEU serta validasi kepakaran manusia untuk menjamin keamanan makna hukum.

1.7 Sistematika Penulisan

Laporan Tugas Akhir ini disusun menjadi lima bab utama yang saling berkaitan satu sama lain. Sistematika penulisan ini dibuat untuk memberikan gambaran umum mengenai isi dari masing-masing bab, urutan pembahasan, serta bagaimana setiap tahapan penelitian terhubung dari awal hingga akhir.

Pada bagian awal, Bab I Pendahuluan berperan sebagai pengantar yang menjelaskan latar belakang masalah, yaitu mengenai adanya kesenjangan literasi hukum di masyarakat. Bab ini juga memuat rumusan masalah, batasan penelitian agar pembahasan tetap fokus, tujuan, manfaat, hingga kerangka berpikir. Secara garis besar, bab ini menjadi fondasi yang menjelaskan alasan mengapa sistem penyederhanaan teks hukum ini sangat penting untuk dikembangkan.

Setelah memahami pokok permasalahannya, Bab II Tinjauan Pustaka dan Landasan Teori membahas berbagai literatur dan penelitian terdahulu yang relevan.

Bab ini menguraikan teori-teori dasar yang mendukung jalannya penelitian, mulai dari konsep *Natural Language Processing* (*NLP*), karakteristik gaya bahasa hukum, arsitektur model IndoT5, hingga metrik evaluasi seperti *SARI* dan *BLEU*. Kajian pustaka ini berfungsi sebagai pijakan teori yang memvalidasi langkah-langkah teknis di bab selanjutnya.

Bermodalkan landasan teori tersebut, Bab III Metodologi Penelitian menjelaskan langkah-langkah kerja yang dilakukan menggunakan kerangka CRISP-DM (*Cross-Industry Standard Process for Data Mining*). Bab ini merinci proses teknis penelitian, mulai dari spesifikasi perangkat keras (GPU) dan perangkat lunak yang dipakai, cara pengumpulan dan pembersihan data, pengaturan *hyperparameter* untuk melatih model, hingga skenario pengujiannya. Semua rancangan di bab ini menjadi panduan teknis yang siap untuk dieksekusi.

Langkah-langkah teknis tersebut kemudian dipraktikkan dan dibahas hasilnya pada Bab IV Hasil dan Pembahasan. Bab ini merupakan inti dari eksperimen Tugas Akhir. Pembahasannya mencakup hasil analisis data (EDA), proses pelatihan model dari awal hingga mencapai hasil yang stabil, evaluasi skor metrik otomasi, serta hasil penilaian langsung dari pakar hukum dan pakar linguistik. Di bab ini pula, hasil akhir model ditampilkan dalam bentuk antarmuka aplikasi web (*MLOps*). Rangkaian temuan pada bab ini nantinya akan digunakan sebagai bahan untuk menarik kesimpulan.

Terakhir, seluruh proses dan temuan penelitian dirangkum pada Bab V Penutup. Bab ini berisi kesimpulan yang secara spesifik menjawab rumusan masalah dan tujuan yang telah ditetapkan di awal Bab I. Selain itu, bab ini juga memberikan saran bagi para peneliti selanjutnya yang ingin mengembangkan atau menyempurnakan sistem penyederhanaan teks hukum di masa mendatang.