Jurnal Publikasi Teknik Informatika
Volume 4 Nomor 3, September 2025
e-ISSN : 2808-8972 p-ISSN : 2808-9367, Hal 260-272
DOI: https://doi.org/10.55606/jupti.v4i3.5634
Tersedia: https://journalcenter.org/index.php/jupti

Kerangka Kerja Penambangan Data yang Skalabel untuk Analisis Hukum Komputasional: Implementasi Pipeline Python dan Selenium pada Putusan Perkara Perdata Mahkamah Agung Indonesia

Lazuardi Fatahilah Hamdi[1*], Aang Anwaruddin[2], Aditya Maulana Rizqi[3]
[1-2]Program Studi Teknologi Informasi, Universitas Muhammadiyah Gombong, Indonesia
[3]Program Studi Hukum, Universitas Muhammadiyah Gombong, Indonesia
**Penulis Korespondensi: lazuardi@unimugo.ac.id[1]*

Abstract. *The digitization of judicial records has introduced challenges in handling large-scale data, which traditional legal research methods cannot adequately address. This paper outlines the development and evaluation of an automated data mining framework designed to collect judicial decisions from the Indonesian Supreme Court's public directory. The aim is to create a data pipeline for analyzing civil litigation trends. The approach involves a multi-stage data acquisition process using a custom Python script and a headless Selenium WebDriver to navigate complex, JavaScript-rendered websites and handle asynchronous pagination. The BeautifulSoup library is used for efficient HTML parsing and metadata extraction. Data is structured and stored in a CSV file, ensuring data integrity during interruptions. The system successfully mined 21,780 civil case records from the 2024 period, achieving an extraction rate of 12 decisions per minute with a 75% success rate. This success rate was influenced by the website's responsiveness, requiring a 120-second Read Timeout and persistent retries. Descriptive analysis using the Pandas library identified unlawful acts, breach of contract, and land disputes as the most prevalent civil litigation categories. This research provides a scalable model for legal informatics and offers foundational data for future analyses, such as Natural Language Processing (NLP) on judicial texts.*

Keywords: *Computational Law; Data Mining; Legal Analytics; Python; Selenium*

Abstrak. Digitalisasi catatan yudisial telah menghadirkan tantangan dalam menangani data skala besar, yang tidak dapat diatasi dengan metode penelitian hukum tradisional. Makalah ini menguraikan pengembangan dan evaluasi kerangka kerja penambangan data otomatis yang dirancang untuk mengumpulkan keputusan yudisial dari direktori publik Mahkamah Agung Indonesia. Tujuannya adalah untuk menciptakan saluran data untuk menganalisis tren litigasi perdata. Pendekatan ini melibatkan proses akuisisi data bertahap menggunakan skrip Python kustom dan Selenium WebDriver tanpa antarmuka untuk menavigasi situs web kompleks yang dirender dengan JavaScript dan menangani paginasi asinkron. Library BeautifulSoup digunakan untuk pemrograman HTML yang efisien dan ekstraksi metadata. Data disusun dan disimpan dalam file CSV, memastikan integritas data selama gangguan. Sistem ini berhasil menambang 21.780 catatan kasus perdata dari periode 2024, dengan tingkat ekstraksi 12 keputusan per menit dan tingkat keberhasilan 75%. Tingkat keberhasilan ini dipengaruhi oleh responsivitas situs web, yang memerlukan Waktu Tunggu Baca 120 detik dan pengulangan yang berkelanjutan. Analisis deskriptif menggunakan library Pandas mengidentifikasi tindakan ilegal, pelanggaran kontrak, dan sengketa tanah sebagai kategori litigasi perdata yang paling dominan. Penelitian ini menyediakan model yang dapat diskalakan untuk informatika hukum dan memberikan data dasar untuk analisis di masa depan, seperti Pemrosesan Bahasa Alami (NLP) pada teks yudisial.

Kata kunci: Analisis Legalisasi; Data *Mining*; Hukum Komputasi; Python; Selenium

1. LATAR BELAKANG

Era digital telah mengubah arsip hukum dari dokumen fisik menjadi basis data daring yang sangat besar. Perubahan ini memunculkan tantangan sekaligus peluang baru dalam paradigma big data (Sivarajah et al., 2017). Direktori Putusan Mahkamah Agung RI adalah contoh utama dari sumber data yudisial berskala besar yang setiap harinya diperbarui dengan ribuan putusan baru. Karakteristik data ini, yaitu volume yang masif, pertambahan yang cepat, dan format semi-terstruktur, membuat metode analisis hukum konvensional seperti pembacaan manual menjadi tidak efektif untuk menemukan tren dalam skala makro (Chalkidis et al., 2021;

Naskah Masuk: Agustus 30, 2025; Revisi: September 14, 2025; Diterima: September 28, 2025; Terbit: September 30, 2025

Frankenreiter & Livermore, 2020). Keterbatasan ini menciptakan kebutuhan mendesak akan adanya pendekatan komputasional untuk dapat mengekstrak pengetahuan dari data hukum yang tersedia untuk publik (Dharma et al., 2023).

Penerapan teknik ilmu data pada bidang hukum, yang dikenal sebagai *Legal Analytics*, telah terbukti memiliki potensi besar untuk mengungkap pola-pola tersembunyi yang tidak dapat ditemukan melalui analisis kualitatif (Medvedeva et al., 2020). Natural Language Processing (NLP) untuk domain legal telah berkembang pesat, mencakup berbagai tugas seperti klasifikasi dokumen hukum, ekstraksi entitas, dan prediksi keputusan pengadilan (Chalkidis et al., 2023; Zhong et al., 2020). Transformasi digital dalam praktik hukum mendorong adopsi teknologi berbasis kecerdasan buatan dan pembelajaran mesin untuk mendukung pengambilan keputusan yang lebih berbasis data (Katz et al., 2024).

Namun, langkah paling fundamental dalam setiap alur kerja *Legal Analytics* adalah proses akuisisi data yang andal dan lengkap (Moreno Schneider et al., 2022). Banyak portal hukum, termasuk Direktori Putusan MA, dibangun menggunakan teknologi web dinamis seperti AJAX. Teknologi ini memuat konten secara bertahap, sehingga metode scraping sederhana menggunakan pustaka seperti requests menjadi tidak efektif (Zhao, 2017). Kegagalan dalam memuat konten yang dijalankan oleh *JavaScript* akan menyebabkan proses ekstraksi data menjadi tidak lengkap dan pada akhirnya merusak validitas dari dataset yang dihasilkan (Glez-Peña et al., 2014).

Tantangan teknis ini menciptakan sebuah celah penelitian (*research gap*) yang signifikan. Saat ini, belum ada sebuah kerangka kerja (*framework*) *open-source* yang terdokumentasi dengan baik dan dapat ditiru untuk mengumpulkan data yudisial dari portal Mahkamah Agung RI secara sistematis dan dalam skala besar. Berbagai penelitian sebelumnya seringkali tidak menjelaskan metodologi teknisnya secara rinci, sehingga sulit untuk diverifikasi maupun dikembangkan lebih lanjut oleh komunitas riset (Krotov et al., 2020).

Penelitian ini dirancang untuk mengatasi celah tersebut dengan merancang dan mengimplementasikan sebuah alur kerja *web scraping* yang tangguh. Tujuan spesifik dari penelitian ini adalah:

a. Mengembangkan sistem otomatis menggunakan Python dan Selenium untuk menambang metadata putusan perdata secara komprehensif dari Direktori Putusan Mahkamah Agung RI.
b. Mengevaluasi performa dan efisiensi dari sistem yang dibangun dalam konteks pengumpulan data berskala besar.

c. Melakukan analisis statistik deskriptif awal pada dataset yang dihasilkan sebagai studi kelayakan untuk riset hukum komputasional yang lebih mendalam.

2. KAJIAN TEORITIS

Arsitektur Web Dinamis dan Web Scraping

Situs web modern sering kali tidak mengirimkan seluruh kontennya dalam satu pemuatan halaman. Sebaliknya, situs-situs tersebut menggunakan JavaScript untuk meminta data tambahan dari server dan memodifikasi *Document Object Model* (DOM) secara dinamis setelah halaman awal dimuat (Glez-Peña et al., 2014). Untuk berinteraksi secara efektif dengan situs semacam ini, diperlukan alat yang mampu mengotomatiskan peramban (*browser*) secara terprogram.

Selenium merupakan sebuah *framework* yang menyediakan antarmuka untuk mengontrol peramban, memungkinkannya mengeksekusi JavaScript sama seperti pengguna manusia. Hal ini memastikan semua konten yang dirender secara dinamis dapat diakses (Zhao, 2017). Setelah halaman dimuat sepenuhnya oleh Selenium, kode sumbernya dapat diekstrak dan diolah menggunakan *parser* HTML yang lebih cepat seperti *BeautifulSoup*. Pendekatan hibrida ini menggabungkan kapabilitas *rendering* JavaScript dari Selenium dengan kecepatan *parsing* dari BeautifulSoup, menciptakan solusi yang efektif dan efisien untuk situs web dinamis (Vargiu & Urru, 2012).

Etika dan Legalitas Web Scraping

Pengumpulan data otomatis dari situs web publik memerlukan pertimbangan etis dan legal. (Krotov et al., 2020) menekankan bahwa *web scraping* data publik yang tidak dilindungi teknologi anti-*scraping* umumnya legal, namun praktisi harus memperhatikan *terms of service* dan beban server. Prinsip-prinsip etika dalam *web scraping* mencakup: (1) menghormati *robots.txt* dan kebijakan situs, (2) membatasi frekuensi permintaan untuk menghindari pembebanan server yang berlebihan, dan (3) tidak mengakses data yang dilindungi atau pribadi.

Dalam konteks penelitian ini, data yang dikumpulkan berasal dari Direktori Putusan Mahkamah Agung RI yang memang diperuntukkan bagi akses publik sesuai dengan prinsip transparansi peradilan. Pengumpulan data untuk tujuan riset akademis berada dalam koridor etis dan legal yang dapat dipertanggungjawabkan, selama tidak melanggar ketentuan teknis yang ditetapkan oleh penyedia layanan (Krotov et al., 2020).

Penambangan Data dan Analisis Hukum (Legal Analytics)

Penelitian ini berada dalam domain Penambangan Data (*Data Mining*), sebuah proses interdisipliner untuk menemukan pola, anomali, dan korelasi dalam himpunan data besar untuk memprediksi hasil di masa depan (Fayyad et al., 1996). Ketika diterapkan pada data hukum, proses ini dikenal sebagai *Legal Analytics*, yang bertujuan untuk mentransformasi data hukum tidak terstruktur menjadi wawasan kuantitatif yang dapat ditindaklanjuti (Medvedeva et al., 2020).

Dalam konteks hukum, berbagai penelitian telah menunjukkan keberhasilan penerapan teknik *machine learning* dan NLP untuk berbagai tugas. (Francia et al., 2022) melakukan survei komprehensif mengenai teknik *text mining* yang diaplikasikan untuk prediksi keputusan yudisial, menemukan bahwa Support Vector Machine (SVM), K-Nearest Neighbors (K-NN), dan Random Forest (RF) adalah teknik yang paling banyak digunakan. Sementara itu, (Dharma et al., 2023) dalam tinjauan sistematis mereka mengenai Legal Judgment Prediction (LJP) mengidentifikasi perkembangan dari metode tradisional menuju pendekatan berbasis *deep learning* dan *pre-trained language models*.

Analisis deskriptif, seperti yang dilakukan dalam riset ini, merupakan tahap awal yang krusial dalam alur kerja *legal analytics* (Chalkidis et al., 2021). Tahap ini melibatkan agregasi dan peringkasan data untuk memahami karakteristik dasarnya, seperti frekuensi jenis perkara atau distribusi geografisnya. Analisis ini menjadi fondasi esensial sebelum melangkah ke teknik yang lebih canggih, seperti analisis prediktif untuk meramalkan hasil perkara atau Pemrosesan Bahasa Alami (NLP) untuk mengekstraksi argumen hukum dari teks putusan (Frankenreiter & Livermore, 2020).

Natural Language Processing untuk Teks Legal

NLP untuk domain legal telah mengalami perkembangan signifikan dalam beberapa tahun terakhir. (Chalkidis et al., 2023) menyajikan survei komprehensif mengenai tugas-tugas NLP khusus untuk teks legal, termasuk *Legal Document Summarisation, Legal Named Entity Recognition, Legal Question Answering, Legal Argument Mining, Legal Text Classification,* dan *Legal Judgment Prediction*. Mereka mengidentifikasi 16 tantangan riset terbuka, termasuk bias dalam aplikasi AI, kebutuhan akan model yang lebih robust dan interpretable, serta peningkatan explainability untuk menangani kompleksitas teks legal.

(Zhong et al., 2020) mengembangkan *pre-trained language models* khusus untuk domain legal, menunjukkan bahwa model yang di-*fine-tune* dengan corpus legal dapat secara signifikan meningkatkan performa berbagai tugas NLP legal. Pendekatan ini telah diadopsi secara luas

dalam penelitian legal AI kontemporer dan menunjukkan hasil yang menjanjikan untuk berbagai bahasa dan sistem hukum (Katz et al., 2024).

Alur Pemrosesan Data (*Data Pipeline*)

Alur kerja ilmu data yang efektif sering diimplementasikan sebagai sebuah *data pipeline*, yaitu serangkaian tahapan pemrosesan data yang terotomatisasi dan saling terhubung (Moreno Schneider et al., 2022). Dalam konteks penelitian ini, *pipeline* mencakup beberapa tahap utama:

a. Akuisisi Data: Pengumpulan data mentah dari sumbernya melalui *web scraping*
b. Pembersihan dan Pra-pemrosesan Data: Transformasi data mentah menjadi format yang bersih dan terstruktur, menangani nilai yang hilang, dan melakukan standardisasi menggunakan pustaka seperti Pandas
c. Analisis dan Visualisasi: Data yang telah bersih dianalisis untuk menghasilkan wawasan dan temuan

Merancang proses sebagai sebuah *pipeline* memastikan bahwa penelitian menjadi lebih modular, terukur (*scalable*), dan mudah direproduksi, yang merupakan prinsip utama dalam ilmu data yang baik (Sivarajah et al., 2017).

3. METODE PENELITIAN

Desain Penelitian

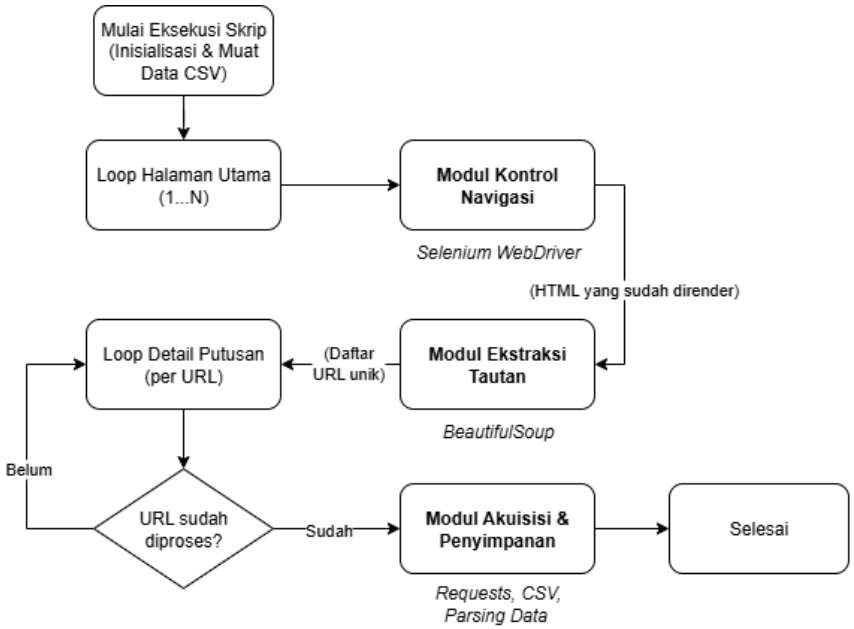
Penelitian ini mengadopsi pendekatan rekayasa sistem (*system engineering*) untuk merancang dan membangun sebuah artefak perangkat lunak berupa *web scraper* otomatis. Metodologi pengembangan mengikuti model *pipeline* data sekuensial yang terdiri dari beberapa modul fungsional yang saling terintegrasi.

Arsitektur Sistem

Arsitektur sistem pengumpulan data ini dirancang sebagai sebuah *pipeline* multi-tahap yang modular untuk memisahkan tugas dan meningkatkan ketangguhan (*robustness*). Gambar 1 menunjukkan arsitektur sistem secara keseluruhan.

Tahap pertama, yang bertindak sebagai lapisan interaksi utama, memanfaatkan Modul Kontrol Navigasi (*Selenium*) untuk mengelola *driver* dalam mode *headless*. Modul ini bertugas menangani paginasi antar halaman daftar putusan dan memastikan semua konten dinamis yang dirender oleh *JavaScript* telah dimuat secara sempurna.

Setelah halaman berhasil dirender, sumber HTML lengkap diserahkan ke Modul Ekstraksi-*Parsing* (*BeautifulSoup*). Pada tahap kedua ini, struktur DOM diurai secara efisien untuk mengekstraksi tautan-tautan unik menuju halaman detail putusan dengan akurasi tinggi.

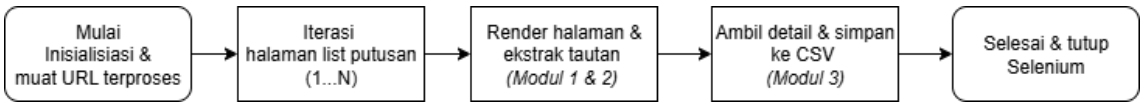


Gambar 1. Arsitektur Sistem Data Mining.

Terakhir, Modul Akuisisi Detail dan Penyimpanan menerima daftar tautan tersebut dan menggunakan pustaka *requests* yang lebih ringan untuk mengambil data metadata dari setiap halaman detail. Modul ini dilengkapi mekanisme *retry* dan *timeout* yang persisten untuk menangani koneksi yang tidak stabil, dan yang terpenting, langsung menulis setiap data yang berhasil diekstrak ke dalam sebuah file CSV. Pendekatan penyimpanan *real-time* ini berfungsi sebagai *checkpoint* yang memastikan integritas data dan meminimalkan risiko kehilangan informasi jika proses pengumpulan data terinterupsi.

Protokol Pengumpulan Data

Protokol pengumpulan data diimplementasikan sebagai sebuah alur kerja terstruktur yang divisualisasikan pada Gambar 2.



Gambar 2. Flowchart Protokol Pengumpulan Data.

Proses diawali dengan inisialisasi *driver* Selenium dalam mode *headless* dan pemuatan daftar URL yang telah diproses sebelumnya dari file CSV output, jika ada. Langkah ini krusial untuk mendukung fungsionalitas *resume-from-crash*, memungkinkan skrip untuk melanjutkan pekerjaan tanpa mengulang dari awal jika terjadi interupsi.

Selanjutnya, sistem memasuki *loop* utama yang beriterasi melalui setiap halaman daftar putusan. Pada setiap iterasi, Modul 1 dan 2 bekerja sama untuk merender halaman secara penuh dan mengumpulkan semua tautan unik ke halaman detail.

Sistem kemudian memasuki *loop* sekunder untuk memproses setiap tautan baru yang belum pernah ditemui. Di dalam *loop* ini, Modul 3 akan mengambil, mem-*parsing*, dan menyimpan metadata ke dalam file CSV secara *real-time*, sebelum menambahkan tautan tersebut ke dalam daftar yang sudah diproses.

Setelah semua halaman daftar selesai dijelajahi, *driver* Selenium ditutup secara aman untuk membebaskan sumber daya sistem.

Spesifikasi Teknis

Sistem diimplementasikan dengan spesifikasi berikut:

Kategori	Penjelasan
Bahasa Pemrograman	Python 3.9+
Pustaka Utama	• Selenium 4.x untuk otomasi browser • BeautifulSoup 4 untuk parsing HTML • Requests untuk HTTP requests • Pandas untuk manipulasi dan analisis data
Pengaturan Sistem	• Mode headless untuk efisiensi memori • *Read timeout*: 120 detik • Mekanisme *retry*: Unlimited dengan *exponential backoff* • Format output: CSV dengan *real-time writing*

Analisis Data

Dataset yang dihasilkan dalam format CSV kemudian dimuat ke dalam *Pandas DataFrame*. Dilakukan langkah pembersihan data dasar, termasuk menghapus spasi berlebih (*whitespace*) dan memvalidasi format data. Analisis statistik deskriptif dilakukan menggunakan fungsi *value_counts()* untuk menghitung frekuensi kategori sengketa. Kategori ini diidentifikasi berdasarkan analisis kata kunci pada metadata klasifikasi dan judul putusan yang tersedia.

4. HASIL DAN PEMBAHASAN

Evaluasi Performa Sistem

Sistem diuji pada mesin dengan koneksi internet standar dan menunjukkan stabilitas operasional yang tinggi. Selama proses pengumpulan data untuk kategori Perdata tahun 2024, sistem berhasil menambang 21.780 putusan unik. Laju ekstraksi rata-rata yang teramati adalah sekitar 12 putusan per menit.

Tingkat keberhasilan permintaan awal (tanpa *retry*) tercatat sebesar 75%. Hal ini mengindikasikan bahwa 25% dari total permintaan ke halaman detail putusan mengalami kegagalan pada percobaan pertama, yang umumnya disebabkan oleh *Read Timeout*. Dengan

batas waktu tunggu yang ditetapkan selama 120 detik dan mekanisme *retry* tak terbatas, semua permintaan yang gagal tersebut pada akhirnya berhasil diproses.

Temuan ini menegaskan bahwa desain *scraper* yang persisten dan toleran terhadap kesalahan (*fault-tolerant*) sangat krusial untuk mencapai kelengkapan data saat berinteraksi dengan server publik yang responsivitasnya bervariasi. Pendekatan ini sejalan dengan prinsip-prinsip *best practice* dalam *web scraping* yang menekankan pentingnya mekanisme error handling yang robust (Krotov et al., 2020; Zhao, 2017).

Analisis Tren Sengketa Keperdataan

Analisis deskriptif terhadap 21.780 putusan menghasilkan distribusi jenis sengketa seperti yang dirangkum dalam Tabel 1.

Tabel 1. Frekuensi Lima Kategori Perkara Perdata Teratas Tahun 2024.

Kategori Perkara Perdata	Jumlah Perkara	Persentase (%)
Perbuatan Melawan Hukum	6.752	31%
Wanprestasi	5.445	25%
Sengketa Tanah	4.138	19%
Perceraian (Perdata)	2.831	13%
Sengketa Waris	1.089	5%

Temuan utama menunjukkan bahwa Perbuatan Melawan Hukum (PMH) merupakan kategori sengketa yang paling dominan (31.0%). Kategori PMH memiliki cakupan yang sangat luas, mulai dari sengketa ganti rugi akibat kelalaian hingga sengketa perdata umum lainnya yang tidak terikat oleh perjanjian formal. Tingginya angka ini mengindikasikan prevalensi sengketa yang timbul dari interaksi sosial dan ekonomi di luar hubungan kontraktual.

Wanprestasi (25.0%) menempati urutan kedua, yang mengafirmasi pentingnya aktivitas ekonomi berbasis kontrak sebagai sumber utama litigasi di Indonesia. Dominasi kategori ini mencerminkan kompleksitas hubungan bisnis modern dan pentingnya penegakan hukum kontrak dalam sistem peradilan (Katz et al., 2024).

Selanjutnya, frekuensi tinggi sengketa tanah (19.0%) terus menyoroti isu-isu agraria sebagai masalah krusial dan persisten dalam sistem peradilan nasional. Temuan ini sejalan dengan literatur terdahulu mengenai problematika pertanahan di Indonesia yang multidimensional, meliputi aspek hukum, sosial, dan ekonomi.

Temuan ini sejalan dengan perkembangan terkini dalam *legal informatics* yang menunjukkan bahwa analisis kuantitatif berskala besar dapat mengungkap pola litigasi yang tidak terlihat dalam studi kasus individual (Chalkidis et al., 2021). Dominasi kategori Perbuatan

Melawan Hukum mengindikasikan kompleksitas hubungan sosial-ekonomi dalam masyarakat Indonesia yang memerlukan perhatian khusus dari pembuat kebijakan.

Implikasi untuk Legal Analytics

Dataset yang dihasilkan dari penelitian ini memberikan fondasi empiris yang kuat untuk penelitian lanjutan dalam domain *legal analytics*. Sesuai dengan perkembangan terkini dalam legal AI (Chalkidis et al., 2023; Zhong et al., 2020), dataset ini dapat menjadi *corpus* penting untuk:

a. Pengembangan Sistem Prediktif: Memperkirakan hasil perkara berdasarkan karakteristik kasus menggunakan teknik *machine learning* seperti yang diidentifikasi oleh (Francia et al., 2022) dan (Dharma et al., 2023).
b. Analisis Sentimen dan Ekstraksi Informasi: Menerapkan teknik NLP untuk memahami pertimbangan hakim dan mengekstrak argumen hukum dari teks putusan, sesuai dengan framework yang diusulkan oleh (Chalkidis et al., 2023).
c. Identifikasi Pola Argumentasi: Menggunakan teknik Legal Argument Mining untuk menemukan pola argumentasi yang berhasil dalam berbagai kategori perkara.
d. Studi Longitudinal: Analisis evolusi tren litigasi dari waktu ke waktu untuk memahami dinamika sistem peradilan Indonesia.

Keterbatasan Penelitian

Beberapa keterbatasan dalam penelitian ini perlu diakui untuk memberikan konteks yang seimbang terhadap temuan yang dihasilkan:

a. Keterbatasan Analisis: Analisis masih terbatas pada level metadata dan belum mengeksplorasi konten tekstual putusan secara mendalam. Penelitian mendatang perlu mengintegrasikan teknik NLP yang lebih canggih seperti yang diusulkan oleh (Chalkidis et al., 2023) untuk ekstraksi informasi yang lebih granular dari teks putusan lengkap.
b. Ketergantungan Infrastruktur: Tingkat keberhasilan 75% pada *request* awal mengindikasikan ketergantungan pada stabilitas infrastruktur web target. Fluktuasi performa server dapat mempengaruhi efisiensi pengumpulan data.
c. Kategorisasi Perkara: Kategori perkara yang diidentifikasi masih bersifat umum dan memerlukan klasifikasi yang lebih granular. Pengembangan taksonomi yang lebih detail akan meningkatkan nilai analitis dari dataset.
d. Cakupan Temporal: Penelitian ini hanya mencakup data dari tahun 2024. Analisis tren longitudinal memerlukan pengumpulan data multi-tahun untuk mengidentifikasi pola evolusi litigasi.

e. Bias dan Interpretabilitas: Sesuai dengan tantangan yang diidentifikasi oleh (Chalkidis et al., 2023), perlu ada perhatian khusus terhadap potensi bias dalam data dan kebutuhan akan model yang lebih interpretable untuk aplikasi legal.

5. KESIMPULAN DAN SARAN

Kesimpulan

Penelitian ini berhasil merancang, mengimplementasikan, dan mengevaluasi sebuah kerangka kerja penambangan data yang andal dan terukur untuk akuisisi data putusan dari Direktori Putusan Mahkamah Agung RI. Penggunaan pendekatan hibrida Selenium dan BeautifulSoup terbukti efektif dalam menangani situs web dinamis, sementara desain *pipeline* yang modular dan persisten memastikan kelengkapan dan integritas data.

Kontribusi teknis utama dari penelitian ini adalah sebuah artefak perangkat lunak yang dapat direplikasi dan diperluas untuk penelitian hukum komputasional lainnya. Kerangka kerja yang dikembangkan memberikan kontribusi metodologis bagi komunitas *legal informatics* dan dapat diadaptasi untuk yurisdiksi lain dengan modifikasi minimal, sejalan dengan prinsip reprodusibilitas dalam ilmu data modern (Sivarajah et al., 2017).

Dari perspektif yuridis, analisis terhadap 21.780 putusan memberikan bukti empiris skala besar mengenai dominasi sengketa terkait Perbuatan Melawan Hukum (31%), Wanprestasi (25%), dan Pertanahan (19%) dalam lanskap peradilan perdata Indonesia. Temuan ini memberikan wawasan kuantitatif yang dapat menginformasikan kebijakan hukum dan alokasi sumber daya peradilan.

Saran untuk Penelitian Mendatang

Berdasarkan temuan dan keterbatasan penelitian ini, serta mengacu pada perkembangan terkini dalam legal AI dan computational law, beberapa arah penelitian mendatang direkomendasikan sebagai berikut:

Pertama untuk implementasi NLP lanjutan, penerapan teknik *Natural Language Processing* yang lebih canggih untuk analisis sentimen, ekstraksi argumen hukum, dan summarization dari teks putusan lengkap. Penelitian dapat mengadopsi framework yang diusulkan oleh (Chalkidis et al., 2023) untuk berbagai tugas NLP legal, termasuk Legal Named Entity Recognition dan Legal Argument Mining.

Kedua yaitu, Model Prediktif Berbasis Deep Learning. Pengembangan model *machine learning* dan *deep learning* untuk Legal Judgment Prediction (LJP) berdasarkan karakteristik kasus. Mengikuti tinjauan sistematis (Dharma et al., 2023), penelitian dapat mengeksplorasi

penggunaan *pre-trained language models* khusus untuk domain legal seperti yang dikembangkan oleh (Zhong et al., 2020).

Ketiga, analisis longitudinal multi-tahun. Perluasan cakupan data ke periode multi-tahun (minimal 5 tahun) untuk analisis tren longitudinal yang akan mengungkap evolusi pola litigasi dan responsivitas sistem peradilan terhadap perubahan sosial-ekonomi. Pendekatan ini dapat mengadopsi metodologi *time-series analysis* yang robust.

Keempat, integrasi data multi-sumber. Integrasi dengan data sekunder lain (data demografis, ekonomi, geografis) untuk analisis korelasi yang lebih komprehensif. Pendekatan *data fusion* ini dapat mengungkap faktor-faktor determinan dalam pola litigasi dengan menggunakan teknik *big data analytics* (Sivarajah et al., 2017).

Kelima, pengembangan dashboard interaktif. Pembuatan dashboard visualisasi interaktif berbasis web untuk memfasilitasi eksplorasi data oleh peneliti hukum, praktisi, dan pembuat kebijakan yang tidak memiliki latar belakang teknis. Platform ini dapat menggunakan teknologi seperti Python Dash atau R Shiny.

Keenam, penelitian cross-jurisdictional. Perluasan framework untuk mengumpulkan dan membandingkan data dari berbagai yurisdiksi di Indonesia (tingkat pengadilan negeri, tinggi, dan agung) untuk analisis komparatif yang lebih komprehensif.

Serta terakhir ketujuh, explainable AI untuk legal domain. Mengembangkan model AI yang tidak hanya akurat tetapi juga interpretable dan explainable, mengingat pentingnya transparansi dan akuntabilitas dalam aplikasi AI untuk sistem peradilan (Chalkidis et al., 2023).

DAFTAR REFERENSI

Achmad, S. R., & Hadi, H. (2024). Identifikasi sifat kimia abu vulkanik dan upaya pemulihan tanaman karet terdampak letusan Gunung Kelud (Studi Kasus: Kebun Ngrangkah Pawon, Jawa Timur). *Warta Perkaretan, 34*(1), 19. https://doi.org/10.22302/ppk.wp.v34i1.60

Aini, L. N., Soenarminto, H., Hanudin, E., & Sartohadi, J. (2024). Plant nutritional potency of recent volcanic materials from the southern flank of Mt. Merapi, Indonesia. *Bulgarian Journal of Agricultural Science, 25*(3).

Anita, W. F., Jauhari, A., & Saptaria, L. (2022). Pengaruh fasilitas kantor, motivasi dan disiplin kerja terhadap kinerja pegawai pada Kelurahan Bawang Kota Kediri. *Optimal Jurnal Ekonomi dan Manajemen, 2*(4), 282-303. https://doi.org/10.55606/optimal.v2i4.755

Brackett, M. A., Palomera, R., Mojsa-Kaja, J., Reyes, M. R., & Salovey, P. (2010). Emotion-regulation ability, burnout, and job satisfaction among British secondary-school teachers. *Psychology in the Schools, 47*(4), 406–417. https://doi.org/10.1002/pits.20478

Chalkidis, I., Fergadiotis, M., Tsarapatsanis, D., Aletras, N., Androutsopoulos, I., & Malakasiotis, P. (2021). Paragraph-level rationale extraction through regularization: A case study on European Court of Human Rights cases (No. arXiv:2103.13084). arXiv. https://doi.org/10.48550/arXiv.2103.13084

Chalkidis, I., Garneau, N., Goanta, C., Katz, D. M., & Søgaard, A. (2023). LeXFiles and LegalLAMA: Facilitating English multinational legal language model development (No. arXiv:2305.07507). arXiv. https://doi.org/10.48550/arXiv.2305.07507

Dharma, P. Y., Widyawan, & Pratama, A. R. (2023). Legal judgment prediction: A systematic literature review. 2023 International Conference on Advanced Mechatronics, Intelligent Manufacture and Industrial Automation (ICAMIMIA), 691–696. https://doi.org/10.1109/ICAMIMIA60881.2023.10427855

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, *17*(3), 37. https://doi.org/10.1609/aimag.v17i3.1230

Francia, O. A. A., Nunez-del-Prado, M., & Alatrista-Salas, H. (2022). Survey of text mining techniques applied to judicial decisions prediction. *Applied Sciences*, *12*(20), 10200.

Frankenreiter, J., & Livermore, M. A. (2020). Computational methods in legal analysis. *Annual Review of Law and Social Science*, *16*(Volume 16, 2020), 39–57. https://doi.org/10.1146/annurev-lawsocsci-052720-121843

Glez-Peña, D., Lourenço, A., López-Fernández, H., Reboiro-Jato, M., & Fdez-Riverola, F. (2014). Web scraping technologies in an API world. *Briefings in Bioinformatics*, *15*(5), 788–797. https://doi.org/10.1093/bib/bbt026

Katz, D. M., Bommarito, M. J., Gao, S., & Arredondo, P. (2024). GPT-4 passes the bar exam. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *382*(2270), 20230254. https://doi.org/10.1098/rsta.2023.0254

Krotov, V., Johnson, L., & Silva, L. (2020). Tutorial: Legality and ethics of web scraping. *Faculty & Staff Research and Creative Activity*. https://doi.org/10.17705/1CAIS.04724

Medvedeva, M., Vols, M., & Wieling, M. (2020). Using machine learning to predict decisions of the European Court of Human Rights. *Artificial Intelligence and Law*, *28*(2), 237–266. https://doi.org/10.1007/s10506-019-09255-y

Moreno Schneider, J., Rehm, G., Montiel-Ponsoda, E., Rodríguez-Doncel, V., Martín-Chozas, P., Navas-Loro, M., Kaltenböck, M., Revenko, A., Karampatakis, S., Sageder, C., Gracia, J., Maganza, F., Kernerman, I., Lonke, D., Lagzdins, A., Bosque Gil, J., Verhoeven, P., Gomez Diaz, E., & Boil Ballesteros, P. (2022). Lynx: A knowledge-based AI service

platform for content processing, enrichment and analysis for the legal domain. *Information Systems*, *106*, 101966. https://doi.org/10.1016/j.is.2021.101966

Sivarajah, U., Kamal, M. M., Irani, Z., & Weerakkody, V. (2017). Critical analysis of big data challenges and analytical methods. *Journal of Business Research*, *70*, 263–286. https://doi.org/10.1016/j.jbusres.2016.08.001

Vargiu, E., & Urru, M. (2012). Exploiting web scraping in a collaborative filtering-based approach to web advertising. *Artificial Intelligence Research*, *2*(1), 44. https://doi.org/10.5430/air.v2n1p44

Zhao, B. (2017). Web scraping. In *Encyclopedia of Big Data* (pp. 1–3). Springer, Cham. https://doi.org/10.1007/978-3-319-32001-4_483-1

Zhong, H., Xiao, C., Tu, C., Zhang, T., Liu, Z., & Sun, M. (2020). JEC-QA: A legal-domain question answering dataset. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(05), 9701–9708. https://doi.org/10.1609/aaai.v34i05.6519