

TUGAS 1

Nama dan NIM Anggota 1: Revinda Visma Novatalia dan 202210370311176

Nama dan NIM Anggota 2: Tantri Romadhoni Siswining Ndaru dan 202210370311240

No	Judul Penelitian dan URL artikel	Author	Abstrak	Dataset yang digunakan	Metode-metode IR yang digunakan	Hasil Penelitian
1.	The impact of fine tuning in LLaMA on hallucinations for named entity extraction in legal documents https://arxiv.org/abs/2506.08827v1	Francisco Vargas, Alejandro González Coene, Gaston Escalante, Exequiel Lobón, Manuel Pulido (2025)	Mengekstrak informasi tentang kecelakaan lalu lintas dari dokumen hukum penting untuk menghitung biaya asuransi, tetapi mengekstrak entitas seperti tingkat kecacatan dan kompensasi sulit dilakukan karena argumen yang kompleks dalam putusan hakim. Penelitian ini mengusulkan metode dua tahap: segmentasi dokumen untuk menemukan bagian yang relevan, diikuti dengan ekstraksi entitas.	<ul style="list-style-type: none"> 650 putusan hukum → difilter menjadi 278 dokumen relevan (kasus kecelakaan tanpa korban meninggal) Dibuat dua dataset utama: Dataset 1 = 1120 sampel, Dataset 2 = 861 sampel, serta test set = 30 dokumen (120 sampel) 	Segmentasi teks menggunakan dua pendekatan: <ol style="list-style-type: none"> RegEx (aturan simbol %) Vektorisasi semantik blok 120 token → embedding (MiniLM L12 v2 / text-embedding-ada-002) + pencarian dengan FAISS. Ekstraksi entitas dilakukan melalui RAG + LLM prompting.	<ul style="list-style-type: none"> Metode klasik yang menggunakan ekspresi reguler mencapai akurasi sebesar 39,5%. Model GPT-4 Turbo menghasilkan akurasi tertinggi, yaitu 86,1%. Pada model LLaMA-2 70B, akurasi meningkat dari 61,7% menjadi 79,4% setelah proses fine-tuning. Model LLaMA-3 8B menunjukkan kinerja dasar yang kuat, dengan akurasi 76,6%. Proses penyempurnaan secara drastis mengurangi tingkat

			<p>Teknik segmentasi dibandingkan: ekspresi reguler klasik versus vektorisasi blok teks menggunakan model multibahasa (text-embedding-ada-002/MiniLM-L12-v2).</p> <p>Model bahasa besar (LLaMA-2 7b/70b, LLaMA-3 8b, GPT-4 Turbo) diterapkan dengan prompting; LLaMA disesuaikan menggunakan LoRA.</p> <p>Halusinasi tinggi pada LLaMA-2 7b berkurang drastis setelah penyesuaian.</p> <p>Metode baru ini outperformed metode klasik (akurasi 39,5%), dengan akurasi tertinggi dicapai oleh: LLaMA-2</p>		<p>1. Model LLM: LLaMA-2 (7B, 70B), LLaMA-3 (8B), GPT-3.5, GPT-4 Turbo.</p> <p><i>Fine-tuning</i> pada LLaMA menggunakan QLoRA (quantized LoRA)</p>	<p>halusinasi, terutama pada model kecil seperti 7B.</p> <ul style="list-style-type: none"> • Pendekatan berbasis vektor menghasilkan segmentasi akurat sekitar 83% dan secara signifikan mempercepat ekstraksi entitas hukum.
--	--	--	--	--	---	---

			70B yang disesuaikan (79,4%), LLaMA-3 8B (76,6%), dan GPT-4 Turbo (86,1%).			
2.	LEGAL-BERT: The Muppets straight out of Law School — (LEGAL-BERT, 2020). (https://aclanthology.org/2020.findings-emnlp.261.pdf)	I. Chalkidis et al. (2020)	Memperkenalkan keluarga model BERT yang dilatih khusus pada korpora hukum (legislation, contracts, court cases). Diusulkan hyperparameter dan fine-tuning untuk tugas downstream legal NLP.	<ul style="list-style-type: none"> • Korpus hukum besar (legislation, case law, contracts) yang digunakan untuk pretraining • Legal-BERT; dataset downstream : LexGLUE (berbagai tugas hukum). 	Pretraining BERT pada data domain-hukum; fine-tuning untuk tugas klasifikasi / retrieval (surrogate retrieval via embedding/semantic similarity).	Model domain-spesifik (Legal-BERT) meningkatkan performa pada tugas legal NLP dibanding model umum; menunjukkan keuntungan pretraining domain-khusus untuk representasi teks hukum.
3.	Processing Long Legal Documents with Pre-trained Transformers: Modding LegalBERT and Longformer (2022). https://aclanthology.org	D. Mamakas et al. (2022)	Membahas tantangan dokumen hukum yang panjang dan evaluasi model pretrained (Longformer, BigBird, LegalBERT, dll.)	LexGLUE dan kumpulan dataset tugas-tugas panjang di domain hukum.	Pendekatan untuk teks panjang: Longformer/BigBird (dilatasi perhatian), hierarchical encoding (paragraph → document),	Menemukan bahwa model dengan mekanisme atensi untuk dokumen panjang (Longformer) dan pendekatan hierarchical sering lebih baik pada tugas yang melibatkan konteks dokumen

	g/2022.nllp-1.11.pdf?utm_source=chatgpt.com		pada tugas-tugas panjang (classif., retrieval). Menyarankan arsitektur/strategi untuk teks panjang.		hybrid TF-IDF + hierarchical BERT.	penuh; LegalBERT tetap kuat pada bagian yang dipangkas/fine-tuned
4.	A Hybrid Approach to Information Retrieval and Answer Generation for Regulatory Texts (https://huggingface.co/papers/2502.16767)	Jhon Rayo, Raúl de la Rosa, Mario Garrido (2024)	Teks regulasi yang panjang dan kompleks menimbulkan tantangan signifikan bagi sistem pengambilan informasi (IR) dalam melaksanakan fungsi kepatuhan. Penelitian ini mengusulkan sistem IR hibrida yang menggabungkan pencarian leksikal (BM25) dengan pencarian semantik (melalui transformer yang dilatih) untuk mengekstrak informasi penting dari dataset regulasi yang besar. Model transformer disesuaikan untuk	<ul style="list-style-type: none"> Kumpulan data ObliQA terdiri dari 27.869 pertanyaan regulasi yang diekstraksi dari 40 dokumen yang disediakan oleh Abu Dhabi Global Markets (ADGM). Kumpulan data ini dibagi menjadi tiga bagian: pelatihan (22.295), pengujian (2.786), dan 	Sistem Pencarian Informasi (IR) dikembangkan menggunakan tiga pendekatan utama: <ol style="list-style-type: none"> BM25 sebagai dasar leksikal Semantic Retriever berbasis model BAAI/bge-small-en-v1.5 yang telah disesuaikan. Sistem hibrida yang menggabungkan kedua pendekatan tersebut dengan bobot $\alpha = 0.65$ untuk menyeimbangkan hasil semantik dan leksikal. 	<ol style="list-style-type: none"> Model hibrida mencapai Recall@10 = 0.8333 dan MAP@10 = 0.7016, melampaui baik BM25 maupun retriever semantik tunggal. Model semantik yang disesuaikan menunjukkan peningkatan signifikan dibandingkan model dasar (Recall@10 meningkat dari 0.7017 \rightarrow 0.8111; MAP@10 dari 0.5357 \rightarrow 0.6261). Pada tahap pembangkitan jawaban, kombinasi Hybrid Retriever + GPT-3.5 Turbo memberikan skor

			<p>beradaptasi dengan lingkungan regulasi, dan temuan dari proses ini diterapkan dalam kerangka kerja Retrieval-Augmented Generation (RAG) untuk menghasilkan respons mendalam dengan bantuan Model Bahasa Besar (LLMs). Data eksperimental menunjukkan bahwa pendekatan hibrida ini secara signifikan outperforms metode leksikal atau semantik murni, terutama dalam hal Recall@10 dan MAP@10. Penelitian ini juga menekankan pentingnya mengintegrasikan pencarian berbasis kata dan berbasis makna untuk memperkuat</p>	<p>validasi (2.788), yang disimpan dalam format JSON bersama dengan pasangan pertanyaan–artikel yang relevan.</p>	<p>Tahap pra-pemrosesan meliputi normalisasi teks, penghapusan kata stop, stemming (Snowball), dan tokenisasi unigram dan bigram. Proses pembangkitan jawaban dilakukan menggunakan metode</p> <ol style="list-style-type: none"> 1. Retrieval-Augmented Generation (RAG) dengan model GPT-3.5 Turbo. 2. GPT-4o Mini, dan LLaMA 3.1, yang dievaluasi menggunakan metrik RePASs berdasar XXaspek relevansi, kontradiksi, dan cakupan kewajiban. 	<p>RePASs tertinggi (0.57) dibandingkan dengan GPT-4o Mini (0.44) dan LLaMA 3.1 (0.37).</p> <ol style="list-style-type: none"> 4. Sistem ini berhasil mengurangi kesalahan kontradiksi dan meningkatkan cakupan kewajiban hukum dalam jawaban. 5. Para penulis menekankan potensi besar pendekatan hibrida dan RAG untuk sistem QA berbasis regulasi di masa depan
--	--	--	---	---	--	--

			penerapan pemrosesan bahasa alami (NLP) di sektor regulasi dan kepatuhan.			
5.	From Fact Drafts to Operational Systems: Semantic Search in Legal Decisions Using Fact Drafts (MDPI) (https://doi.org/10.3390/bdcc8120185)	G.M. Csányi (2024)	Fokus pada searching semantik di dokumen keputusan pengadilan Hungaria, dengan “fact drafts” sebagai representasi.	1172 keputusan pengadilan Hungaria	Semantic similarity search	<ul style="list-style-type: none"> • Dari kumpulan 1172 keputusan pengadilan Hungaria, sistem semantic similarity dengan “fact drafts” (ringkasan fakta) diuji untuk menemukan kasus yang sebanding. MDPI+1 • Menunjukkan bahwa representasi fakta hukum (fact drafts) meningkatkan performa pencarian semantik dibanding hanya menggunakan teks mentah. • Temuan penting: sistem ini lebih praktis/operasional meskipun masih menghadapi tantangan evaluasi (cek manual

						<p>dokumen lengkap mahal).</p> <ul style="list-style-type: none"> Menyimpulkan bahwa pemanfaatan ringkasan fakta struktur membantu dalam IR hukum.
6.	<p>LePaRD: A Large-Scale Dataset of Judicial Citations to Precedent</p> <p>(https://huggingface.co/papers/2311.09356)</p>	<p>Robert Mahari (MIT & Harvard Law School), Dominik Stambach (ETH Zurich), Elliott Ash (ETH Zurich), Alex “Sandy” Pentland (MIT) (2023)</p>	<p>Studi ini memperkenalkan LePaRD (Legal Passage Retrieval Dataset), sebuah dataset berskala besar yang terdiri dari jutaan contoh referensi hakim federal AS terhadap preseden hukum. LePaRD dirancang untuk memfasilitasi penelitian di bidang pengambilan teks hukum, yaitu proses mengidentifikasi bagian-bagian relevan dari putusan pengadilan sebelumnya sambil mempertimbangkan konteks argumen hukum. Kumpulan data ini disusun dari Proyek Akses</p>	<p>Kumpulan Data LePaRD, yang dibangun dari Proyek Akses Putusan Hukum (CAP), berisi</p> <ul style="list-style-type: none"> 1,7 juta putusan hukum federal AS (3,9 juta pasangan konteks-kutipan unik). Berisi 1,7 juta putusan hukum federal AS (3,9 juta pasangan konteks-kutipan unik). 	<ul style="list-style-type: none"> BM25 (Pencarian Jarang) dengan Anserini. SBERT (Pencarian Padat). SBERT yang Disesuaikan (Kerugian Ganda Negatif). LEGAL-BERT (Klasifikasi Berbasis Domain) DistilBERT Klasifikasi untuk klasifikasi terawasi. 	<ul style="list-style-type: none"> Pendekatan yang menggunakan klasifikasi, khususnya model DistilBERT, menghasilkan kinerja terbaik dengan recall@10 mencapai 81,2% pada subset data 10K. BM25 menunjukkan kinerja yang kurang memuaskan akibat keterbatasan dalam kesamaan leksikal. Model SBERT yang telah disesuaikan (fine-tuned) mengalami peningkatan signifikan dibandingkan dengan versi generiknya. Temuan

			Putusan Hukum (CAP) dan mencakup jutaan putusan hukum federal yang telah diterbitkan. Selain itu, penelitian ini mengevaluasi berbagai metode pengambilan informasi (IR) dan menyimpulkan bahwa pendekatan pengambilan berbasis klasifikasi menghasilkan kinerja terbaik, meskipun tingkat recall maksimum hanya mencapai 38%, menunjukkan kompleksitas tinggi dalam pengambilan preseden hukum			ini menegaskan bahwa pengambilan data hukum memerlukan metode yang lebih canggih daripada sekadar pencocokan kata kunci.
7.	Learning Interpretable Legal Case Retrieval via Knowledge-Guided Case Reformulation (https://arxiv.org/pdf/2406.19760)	C. Deng dkk. (2024)	Mengusulkan metode reformulasi kasus hukum berbasis pengetahuan agar retrieval menjadi lebih interpretable dan efektif.	Benchmark retrieval kasus hukum (dua dataset)	Reformulation + Retrieval (LLM guided)	<ul style="list-style-type: none"> Mengusulkan pendekatan “KELLER” yang menggunakan pengetahuan hukum (misalnya jenis tindak pidana + pasal) untuk mereformulasi

						<p>kasus hukum menjadi fakta-sub kasus agar retrieval lebih tepat & interpretable.</p> <ul style="list-style-type: none"> • Eksperimen pada dua benchmark legal case retrieval menunjukkan bahwa KELLER outperform metode sebelumnya dalam performa retrieval & juga lebih robust pada query kasus kompleks. • Implikasi: integrasi domain-knowledge (legal expert knowledge) ke dalam model retrieval hukum meningkatkan relevansi dan interpretabilitas sistem IR hukum.
8.	Challenges and Considerations in Annotating Legal Data: Comprehensive Overview	Harshil Darji, Jelena Mitrovic, Michael Granitzer (2024)	Studi ini mengkaji berbagai tantangan dan aspek yang perlu dipertimbangkan dalam proses anotasi data hukum,	<ul style="list-style-type: none"> • Open Legal Data (Jerman) berisi 43.337 kasus hukum (1,1 	<ul style="list-style-type: none"> • Ekstraksi informasi dilakukan menggunakan ekspresi reguler dan model BERT yang 	<ul style="list-style-type: none"> • Model BERT yang telah melalui proses fine-tuning mampu mengidentifikasi 19 entitas hukum, seperti hakim, pengacara, undang-

	<p>(https://huggingface.co/papers/2407.17503)</p>		<p>mulai dari tahap pemilihan dataset awal hingga pengembangan pedoman anotasi yang seragam. Faktor-faktor krusial seperti kompleksitas bahasa hukum, variasi struktur dokumen, dan kebutuhan partisipasi ahli hukum memainkan peran penting dalam memastikan akurasi dan konsistensi data. Studi ini juga menyoroti hambatan teknis, termasuk ekstraksi teks dari format HTML yang tidak konsisten, serta penggunaan alat anotasi seperti Microsoft Word, yang mempersulit transformasi ke format terstruktur (seperti CoNLL). Penulis</p>	<p>GB) dengan elemen seperti <i>tenor</i>, <i>tatbestand</i>, <i>gründe</i>, dll.</p> <ul style="list-style-type: none"> • Legal-Entity-Recognition dataset untuk pelatihan model NER. • Dataset anotasi hukum baru (2.944 referensi hukum, 21 label) dibuat bersama tiga ahli hukum dan tersedia di Hugging Face. • Dataset kebijakan privasi beranotasi entitas GDPR dari 45 situs web 	<p>telah disesuaikan untuk pengenalan 19 entitas hukum (akurasi F1 = 99,29).</p> <ul style="list-style-type: none"> • Anotasi manual dilakukan dengan bantuan ahli hukum menggunakan Microsoft Word, kemudian data diekstraksi menggunakan ekspresi XPath agar dapat dikonversi ke format CoNLL. • Pembersihan dan pengorganisasian data hukum dilakukan dengan menyesuaikan tag HTML (<h2>) 	<p>undang, peraturan, kontrak, dan sebagainya, dengan skor F1 sebesar 99,29.</p> <ul style="list-style-type: none"> • Tingkat kesepakatan di antara para anotator, yang diukur menggunakan Fleiss' Kappa sebesar 0,87, menunjukkan konsistensi yang sangat baik. • Pendekatan XPath berhasil mengekstrak semua komentar dari dokumen Word, yaitu 100%, berbeda dengan alat seperti Aspose dan python-docx yang hanya mampu mendapatkan sekitar 35%. • Data set yang dihasilkan tersedia secara publik di platform Zenodo dan Hugging Face, mendukung
--	--	--	---	---	--	---

			mengusulkan pendekatan teknis yang melibatkan XPath dan model BERT yang disesuaikan untuk identifikasi entitas hukum (NER) dengan tingkat akurasi yang tinggi.	(mis. Amazon, Airbnb).		pengembangan penelitian lebih lanjut di bidang pemrosesan bahasa alami untuk konteks hukum di Eropa.
9.	EurLexGLUE: A Benchmark Dataset for Legal Language Processing in the EU (https://arxiv.org/abs/2110.00976v4)	Ilias Chalkidis, Sophie Niklaus, Nils Holzenberger, Abhik Jana, Tommaso Caselli, Andreas Vlachos, Ion Androutsopoulos (2025)	Studi ini memperkenalkan EurLexGLUE, sebuah benchmark untuk mengevaluasi kemampuan model bahasa dalam memahami teks hukum multibahasa Uni Eropa. Kumpulan data ini mencakup delapan tugas utama, seperti klasifikasi dokumen, ekstraksi argumen, dan deteksi implikasi, dengan sumber data dari EUR-Lex dan Identifier Legislasi Eropa (ELI) dalam 24 bahasa. EurLexGLUE telah	<ul style="list-style-type: none"> Kumpulan data EurLexGLUE, mencakup 8 subtask pemrosesan bahasa alami (NLP) hukum, termasuk klasifikasi multi-label, transfer tanpa pelatihan (zero-shot transfer), dan deteksi implikasi multibahasa. 	<ul style="list-style-type: none"> Penerapan berbagai Model Bahasa Hukum seperti Legal-BERT, mBERT, XLM-RoBERTa, dan LEGAL-XLM-R. Eksperimen pengambilan informasi menggunakan BM25 dan model pengambilan padat berbasis transformer untuk pencocokan dokumen. Evaluasi mencakup 	<ul style="list-style-type: none"> Model yang digunakan: Legal-BERT, mBERT, XLM-RoBERTa, dan LEGAL-XLM-R. Eksperimen melibatkan: <ol style="list-style-type: none"> BM25 (pencarian leksikal). Model dense retrieval berbasis transformer untuk pencocokan dokumen. Evaluasi mencakup:

			<p>menjadi standar evaluasi untuk model hukum multibahasa seperti Legal-BERT, mBERT, dan XLM-RoBERTa.</p>	<ul style="list-style-type: none"> • Sumber data: EUR-Lex dan ELI, yang berisi lebih dari 65 juta segmen teks hukum UE dalam 24 bahasa resmi. • Data tambahan: JRC-Acquis dan taksonomi EuroVoc untuk penandaan konsep hukum. 	<p>transfer lintas bahasa dan pembelajaran tanpa data (zero-shot learning) di 24 bahasa.</p>	<ol style="list-style-type: none"> 1. Transfer lintas bahasa (cross-lingual transfer). 2. Pembelajaran tanpa data (zero-shot learning) di 24 bahasa. <ul style="list-style-type: none"> • Legal-XLM-R menunjukkan kinerja terbaik di semua tugas (macro-F1 = 77,8). • Legal-BERT unggul pada tugas monolingual (Bahasa Inggris) dengan F1 = 82,3. • Model multibahasa cenderung menurun performanya saat diterapkan lintas bahasa hukum. • EurLexGLUE kini menjadi benchmark standar untuk penelitian NLP hukum multibahasa di Uni Eropa.
--	--	--	---	---	--	---

10.	<p>Optimizing Legal Text Summarization Through Dynamic Retrieval-Augmented Generation and Domain-Specific Adaptation</p> <p>https://doi.org/10.3390/sym17050633</p>	S Ajay Mukund dkk. (2025)	<p>Membahas legal IR dan summarization dalam dokumen hukum yang besar, membandingkan metode tradisional (Boolean, TF-IDF) vs neural.</p>	Korpa dokumen hukum besar (legislation, case law)	Traditional IR + Neural summarization	<ul style="list-style-type: none"> • Menemukan bahwa pendekatan Dynamic Legal RAG yang menggabungkan retrieval domain-khusus + summarisation model (fine-tuned LLaMA 3.1 8B) berhasil meningkatkan akurasi ringkasan legal dibanding model umum. • Sistem menggunakan BM25 + chunking + NER untuk mengambil konteks hukum spesifik (statutes, precedent). • Catatan: “hallucination” (isi yang tidak berdasar) berkurang ketika menggunakan retrieval domain-legal dibanding model general. • Hasilnya menunjukkan
-----	---	---------------------------	--	---	---------------------------------------	--

						bahwa adaptasi domain dan integrasi retrieval sangat membantu dalam konteks dokumen hukum panjang/kompleks
--	--	--	--	--	--	--