(SISTEM QUESTION ANSWERING BERBASIS TRANSFORMER UNTUK DOKUMEN PDF BERBAHASA INDONESIA MENGGUNAKAN FINE-TUNING INDOBERT-LITE PADA DATASET INDOQA)

TUGAS AKHIR

Diajukan sebagai syarat menyelesaikan jenjang strata Satu (S-1) di Program Studi Teknik Informatika, Fakultas Teknologi Industri, Institut Teknologi Sumatera

Oleh:

Nama Mahasiswa

123123123

PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS TEKNOLOGI INDUSTRI
INSTITUT TEKNOLOGI SUMATERA
LAMPUNG SELATAN

2025

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

DAFTAR RUMUS

DAFTAR KODE

BAB I
PENDAHULUAN

1.1 Latar Belakang

Organisasi dan institusi, termasuk perguruan tinggi, instansi pemerintahan, serta perusahaan bergantung pada dokumen operasional berformat PDF, seperti Standar Operasional Prosedur (SOP), peraturan, buku petunjuk kerja, dan buku pedoman, untuk menjamin konsistensi proses, kepatuhan regulasi, serta transfer pengetahuan. Namun, skala dan kompleksitas dokumen tersebut (banyak bab, istilah teknis, dan rujukan silang) menyulitkan pengguna ketika harus menelusuri jawaban spesifik secara cepat dan tepat; pencarian manual memakan waktu dan rawan salah tafsir karena informasi sering tersebar di beberapa bagian yang saling terkait.

Pendekatan pencarian berbasis kata kunci yang lazim digunakan pada PDF memiliki keterbatasan karena hanya mengandalkan kecocokan leksikal. Pendekatan ini tidak mampu menangkap sinonimi, parafrasa, dan hubungan semantik antarkalimat, serta kurang efektif ketika pertanyaan menuntut pemahaman konteks lintas paragraf atau subbab. Kesenjangan ini menegaskan kebutuhan akan solusi yang melampaui sekadar pencarian kata, yakni solusi yang benar-benar memahami isi dokumen dan maksud pertanyaan pengguna.

Kemajuan Pemrosesan Bahasa Alami (Natural Language Processing/NLP) berbasis arsitektur *transformer* membuka peluang untuk membangun sistem *question answering* (QA) yang mampu memahami pertanyaan dan mengekstraksi jawaban dari konteks dokumen panjang secara lebih presisi [1], [2]. Di ranah bahasa Indonesia, keluarga model IndoBERT, termasuk varian ringan IndoBERT-Lite, menyediakan fondasi efisien untuk diadaptasi (*fine-tuning*) ke tugas QA sehingga tetap terjangkau dari sisi komputasi [3]. Selain itu, integrasi teknik pengambilan informasi (*retrieval*) sebelum

inferensi, sebagaimana ditunjukkan pada paradigma *retrieval-augmented*, dapat membantu sistem menjawab pertanyaan berbasis pengetahuan yang tersebar di dokumen besar [4], sementara studi QA atas dokumen panjang/terstruktur terus menyoroti tantangan praktis terkait segmentasi konteks, pemilihan paragraf relevan, dan keterlacakan sumber (mis. PDF/long-document QA).

Berdasarkan tinjauan tersebut, penelitian ini mengangkat dua *research gap*: (1) belum ada penelitian yang secara khusus mengevaluasi IndoBERT-Lite untuk QA pada dokumen PDF nyata berbahasa Indonesia; (2) belum ada integrasi dan evaluasi sistematis pipeline *retrieval* + QA yang ditujukan untuk dokumen non-dataset (PDF) institusional. Untuk menjawab gap ini, penelitian memanfaatkan model pralatih `Wikidepia/indobert-lite-squad` sebagai titik awal *fine-tuning* dan dataset `jakartaresearch/indoqa` sebagai sumber pasangan konteks–pertanyaan–jawaban berbahasa Indonesia, dengan fokus evaluasi pada skenario dunia nyata.

1.2 Rumusan Masalah

Berdasarkan latar belakang, penelitian ini difokuskan pada perancangan dan evaluasi sistem QA berbahasa Indonesia untuk dokumen PDF institusional dalam skala yang realistis untuk diselesaikan. Rumusan masalah yang dikaji adalah:

RQ1: Bagaimana mengembangkan pipeline QA berbasis *retrieval* yang menerima input PDF + pertanyaan, mengekstraksi dan menyegmentasi teks, melakukan *indexing* serta pemilihan paragraf relevan, dan menghasilkan jawaban beserta referensi paragraf sumber sebagai keluaran yang terstandar?

RQ2: Bagaimana performa model IndoBERT-Lite yang di-*fine-tune* pada dataset IndoQA (diukur dengan metrik EM/F1) ketika digunakan sebagai penjawab ekstraktif pada konteks paragraf hasil *retrieval*, dibandingkan performa pralatihnya?

RQ3: Seberapa akurat sistem end-to-end dalam menjawab pertanyaan berbasis dokumen PDF nyata (mis. SOP atau pedoman) yang dipilih, diukur dengan EM/F1 per pertanyaan, serta apa sumber kesalahan utama dari tahap *retrieval* maupun ekstraksi jawaban?

Rumusan masalah ini akan dijawab melalui rancangan sistem, implementasi perangkat lunak, eksperimen terukur, dan analisis yang dilaporkan sampai **??**.

1.3 Tujuan Penelitian

Merujuk pada Latar Belakang dan Rumusan Masalah, tujuan penelitian ini dirumuskan sebagai berikut:

RO1: Mengembangkan pipeline QA berbasis *retrieval* untuk dokumen PDF berbahasa Indonesia, mencakup ekstraksi/segmentasi PDF, *indexing*, pemilihan paragraf relevan, serta penyajian jawaban berikut referensi paragraf sumber.

RO2: Memperoleh model QA IndoBERT-Lite hasil *fine-tuning* pada dataset IndoQA, dan mengukur peningkatan kinerjanya (EM/F1) dibandingkan model pralatih.

RO3: Mengevaluasi performa sistem QA end-to-end pada dokumen PDF target (mis. SOP atau pedoman), diukur dengan EM/F1 per pertanyaan serta analisis kesalahan dari tahap *retrieval* dan ekstraksi jawaban.

1.4 Batasan Masalah

Batasan yang dimaksud disini ialah batasan dari penelitian tugas akhir yang dilakukan. Batasan masalah ditujukan agar tugas akhir yang dilakukan tidak terlalu luas, dan menjadi lebih realistis untuk diselesaikan.

1. Dokumen yang diuji adalah PDF berbasis teks (bukan hasil pemindaian tanpa OCR), sehingga ekstraksi teks tidak membahas pipeline OCR.
2. Task QA bersifat ekstraktif (jawaban diambil langsung dari paragraf

konteks) dan tidak mencakup generatif bebas.

3. Model utama yang digunakan adalah IndoBERT-Lite; model lain hanya sebagai pembanding ringan jika diperlukan.
4. Data *fine-tuning* utama adalah dataset IndoQA; penyesuaian domain hanya sebatas format PDF target, tanpa kurasi dataset baru berskala besar.
5. Evaluasi kinerja menggunakan metrik EM dan F1 untuk level jawaban, serta analisis kesalahan pada tahap *retrieval* dan ekstraksi jawaban.

1.5 Manfaat Penelitian

Manfaat yang ingin dicapai, diharapkan dapat memberikan dampak positif terhadap mahasiswa, program studi Teknik Informatika, ITERA, pengguna institusional, maupun komunitas akademik/peneliti, ialah sebagai berikut:

1. Tersedianya pipeline QA PDF berbahasa Indonesia yang dapat diterapkan pada dokumen institusional (mis. SOP, peraturan kampus/perusahaan), sehingga akses informasi menjadi lebih cepat dan presisi.
2. Model QA berbasis IndoBERT-Lite yang telah di-*fine-tune* dan tervalidasi, memberikan referensi performa (EM/F1) untuk riset dan pengembangan lebih lanjut.
3. Pemanfaatan dataset internal PDF untuk QA yang selama ini jarang dieksplorasi, membuka peluang studi lanjutan tentang domain adaption dan evaluasi QA pada dokumen dunia nyata.
4. Bagi mahasiswa/program studi: pengalaman merancang, mengimplementasikan, dan mengevaluasi sistem AI terapan dengan batasan sumber daya; bagi akademisi/peneliti: kontribusi studi kasus QA PDF berbahasa Indonesia yang dapat direplikasi atau diperluas.

1.6 Sistematika Penulisan

Sistematika penulisan berisi pembahasan apa yang akan ditulis disetiap Bab. Sistematika pada umumnya berupa paragraf yang setiap paragraf

mencerminkan bahasan setiap Bab.

Bab I

Bab ini berisikan penjelasan latar belakang dari topik penelitian yang berlangsung, rumusan masalah dari masalah yang dihadapi pada penjelasan di latar belakang, tujuan dari penelitian, batasan dari penelitian, manfaat dari hasil penelitian, dan sistematika penulisan tugas akhir.

Bab II

Bab ini membahas mengenai tinjauan pustaka dari penelitan terdahulu dan dasar teori yang berkaitan dengan penelitian ini.

Bab III

Bab ini berisikan penjelasan alur kerja sistem, alat dan data yang digunakan, metode yang digunakan, dan rancangan pengujian.

Bab IV

Bab ini membahas hasil implementasi dan pengujian dari penelitian yang dilakukan, serta analisis dan evaluasi yang dapat dipetik dari hasil.

Bab V

Bab ini membahas kesimpulan dari hasil penelitian dan juga saran untuk penelitian selanjutnya.

BAB II
TINJAUAN PUSTAKA

2.1 Tinjauan Pustaka

Bab ini merangkum literatur utama terkait QA dokumen PDF berbahasa Indonesia, dengan fokus pada (i) arsitektur *transformer*, (ii) *retrieval-augmented* QA, dan (iii) ekosistem model/data Indonesia. Lima karya paling relevan dipilih karena langsung menyangga rancangan pipeline PDF $\rightarrow$ segmentasi $\rightarrow$ retrieval $\rightarrow$ reader IndoBERT-Lite:

1. Saad-Falcon et al. (PDFTriage) [5]: QA dokumen panjang berstruktur (PDF), menekankan segmentasi blok/layout sebelum retrieval.
2. Lovenia et al. (SEACrowd) [6]: ekosistem data/benchmark Asia Tenggara; relevan untuk legitimasi evaluasi bahasa Indonesia.
3. Wilie et al. (IndoNLU/IndoBERT) [7]: model/dataset kunci untuk NLU Indonesia; landasan pemilihan IndoBERT/IndoBERT-Lite.
4. Karpukhin et al. (DPR) [8]: baseline *dense* retrieval untuk memilih paragraf relevan sebelum modul QA.
5. Jurafsky & Martin (SLP 3rd draft) [9]: kerangka teori NLP, IR, dan evaluasi yang netral akademik.

Karya fondasional [1], [2], [3], [4] digunakan sebagai rujukan arsitektur/pipeline pendukung.

2.2 Dasar Teori

Transformer dan self-attention. Vaswani et al. [1] memperkenalkan *multi-head self-attention* yang menggantikan rekuren, memungkinkan pemodelan dependensi jarak jauh secara paralel. Devlin et al. [2], [3] menerapkannya pada BERT/IndoBERT untuk tugas ekstraktif.

Retrieval-augmented QA. Lewis et al. [4] memisahkan tahap *retriever*

(dense) dan reader/generator, sehingga konteks panjang tidak perlu dimuat sekaligus. DPR [8] menjadi baseline *dense retriever* untuk memilih paragraf relevan.

Dokumen panjang/berlayout. PDFTriage [5] menegaskan bahwa segmentasi blok teks/layout pada PDF sebelum retrieval meningkatkan presisi dan keterlacakan paragraf sumber.

Ekosistem Indonesia. IndoNLU/IndoBERT [3], [7] menyediakan backbone dan tolok ukur; SEACrowd [6] memperkuat relevansi evaluasi lintas bahasa Asia Tenggara. SLP [9] memberi kerangka konseptual NLP/IR dan metrik (EM/F1).

2.3 Sintesis Literatur Primer

1. **PDFTriage** [5]: menunjukkan pentingnya segmentasi layout-aware untuk QA PDF; relevan untuk SOP/pedoman multi-bab.
2. **SEACrowd** [6]: menyoroti ketersediaan data dan pentingnya evaluasi bahasa Indonesia dalam lanskap Asia Tenggara.
3. **IndoNLU/IndoBERT(-Lite)** [3], [7]: menyediakan model dasar efisien untuk *fine-tuning* QA Indonesia.
4. **DPR** [8]: *dense retriever* untuk memilih paragraf; cocok digabung dengan reader IndoBERT-Lite.
5. **SLP** [9]: kerangka teori NLP/IR untuk merapikan definisi dan metrik evaluasi.

2.4 State-of-the-Art dan Celah Riset

Praktik mutakhir. Pipeline QA modern memadukan retriever padat (DPR/ColBERT) dan reader transformer; RAG [4] mengurangi beban konteks panjang dengan retrieval terpisah; PDFTriage menambah perspektif layout-aware.

Terdapat beberapa *research gap* yang bisa diidentifikasi:

1. Belum ada evaluasi terukur IndoBERT-Lite untuk QA PDF institusional berbahasa Indonesia dengan struktur bab/subbab.
2. Belum ada perbandingan sistematis segmentasi PDF (blok/halaman vs paragraf logis) terhadap kualitas retrieval dan jawaban.
3. Pelacakan sumber (paragraf/halaman) pada jawaban QA PDF Indonesia belum distandarkan.

2.5 Implikasi untuk Penelitian Ini

1. Merancang pipeline: ekstraksi + segmentasi PDF (layout-aware), indexing (sparse/dense), seleksi paragraf, lalu reader IndoBERT-Lite yang di-*fine- tune*.
2. Menyediakan evaluasi EM/F1 dan analisis kesalahan terpisah antara kegagalan retrieval dan ekstraksi; menyertakan referensi paragraf/halaman.
3. Membandingkan skema segmentasi (layout-aware vs paragraf polos) untuk mengukur dampaknya pada presisi retrieval dan kualitas jawaban.

2.6 Ringkasan Literatur

Tabel 2.1 Ringkasan literatur kunci

No.	Judul/Referensi	Masalah	Metode	Hasil/Relevansi
1	PDFTriage [5]	QA dokumen PDF panjang/berstruktur	Segmentasi layout + QA	Akurasi meningkat saat struktur halaman dipertahankan

No.	Judul/Referensi	Masalah	Metode	Hasil/Relevansi
2	SEACrowd [6]	Kesenjangan data/benchmark SEA	Kumpulan korpus & tugas multimodal	Legitimasi evaluasi bahasa Indonesia
3	IndoNLU/IndoBERT [3], [7]	Kurangnya backbone bahasa Indonesia	Pretrained LM + benchmark	Backbone efisien untuk QA Indonesia
4	DPR [8]	Pemilihan paragraf relevan QA	*Dense passage retrieval*	Baseline retriever untuk pipeline PDF
5	SLP [9]	Kerangka teori NLP/IR	Buku teks	Definisi formal, metrik EM/F1

BAB III
METODE PENELITIAN

3.1 Alur Penelitian

Penelitian ini merancang dan mengevaluasi sistem QA PDF berbahasa Indonesia berbasis *transformer* dengan *fine-tuning* IndoBERT-Lite pada IndoQA [3], [10]. Kerangka kerja mengikuti tahapan:

1. Akuisisi dan kurasi data (IndoQA + PDF institusional).
2. Ekstraksi & segmentasi PDF (layout-aware dan *fixed window*).
3. Indeks & *retrieval* (BM25 baseline; opsional dense).
4. *Reader* ekstraktif (IndoBERT-Lite di-*fine-tune*).
5. Evaluasi berlapis (komponen, end-to-end, ablation, V&V).

3.2 Metode Pengumpulan Data

Sumber data sekunder. Dataset IndoQA [10] (konteks–pertanyaan–jawaban) digunakan untuk *fine-tuning* dan validasi reader, dengan split train/val/test (80/10/10 atau sesuai kartu dataset).

Sumber data dari file atau dokumen berupa PDF. Dipilih 1–2 PDF institusional (SOP/peraturan) berteks (non-scan), tebal $\geq$30 halaman, terstruktur (bab/subbab). Ekstraksi teks memakai `pdfminer.six`/`PyPDF2`; OCR dikecualikan sesuai batasan.

Gold standard PDF. Disusun 100–200 pasang Q–A dari PDF oleh dua anotator; kesepakatan dihitung dengan Cohen's κ (target $\geq$ 0,75). Pertanyaan disusun berdasar kebutuhan domain; data sensitif dihapus, memakai dokumen publik/izin internal dan atribusi sumber.

3.3 Metode Perancangan/Pengembangan

3.3.1 Arsitektur Sistem

Modular: ingestion PDF → segmentasi → indeks → retrieval → reader (QA) → pelacakan sumber → antarmuka. Konsep *retrieval-augmented* [4], [8] dan *transformer* [1], [2], [3] menjadi dasar, sementara PDFTriage menekankan struktur dokumen [5].

3.3.2 Pra-pemrosesan dan Segmentasi

- Normalisasi karakter (Unicode, spasi), penggabungan tanda baca per kalimat/ paragraf. Untuk retrieval sparse: *lowercase*, normalisasi angka/tanggal, *stopword* opsional; untuk reader: teks dipertahankan apa adanya.
- Segmentasi layout-aware: deteksi heading/subheading, pecah per subbagian atau paragraf logis (150–250 kata). Fallback: jendela 384 token dengan *stride* 128. Struktur membantu keterlacakan sumber [5].

3.3.3 Indeks dan Retrieval

- BM25 (Pyserini/Lucene) sebagai baseline: $k1 \approx 0{,}9$–$1{,}2$, $b \approx 0{,}4$–$0{,}75$, *top-k* awal = 10.
- Opsional dense retrieval (mis. bi-encoder SEA-LION) untuk studi banding Recall@k.
- *Top-k* (5–10) paragraf kandidat dikirim ke reader.

3.3.4 Model QA dan Fine-tuning

- Basis: `Wikidepia/indobert-lite-squad` (extractive QA).
- Data: IndoQA train/val/test.
- Hiperparameter awal: `max_seq_length`=384, `doc_stride`=128, `max_answer_length`=30, `batch_size`=16, `lr`=2e-5, `epochs`=3–5, `weight_decay`=0.01, `warmup_ratio`=0.1, fp16, `seed`=42, *early stopping* berdasar F1 (patience 2).

3.3.5 Inferensi End-to-End

Menerima pertanyaan → *retrieve top-k* paragraf → reader memproduksi span jawaban + skor → pilih skor tertinggi → kembalikan jawaban beserta paragraf/halaman sumber.

3.3.6 Kontrol Versi & Reproduksibilitas

- Berkas lingkungan (`requirements.txt`), konfigurasi YAML untuk hiperparameter, *random seed*, dan jejak *commit* model.
- *Experiment tracking* (MLflow/Weights & Biases) untuk metrik.

3.4 Metode Pengujian/Validasi

3.4.1 Evaluasi Komponen

- Retrieval-only: Recall@k (k $\in \{1, 3, 5, 10\}$), MRR@k terhadap paragraf emas.
- Reader-only (IndoQA): EM dan F1 pada val/test IndoQA.

3.4.2 Evaluasi End-to-End PDF

- Dataset: himpunan Q–A anotasi PDF.
- Prosedur: retrieval → reader per pertanyaan; ukur EM/F1; catat paragraf/halaman sumber dan klasifikasi kesalahan (retrieval vs ekstraksi).

3.4.3 Uji Statistik dan Ablasi

- Bootstrap 1000× untuk CI 95% EM/F1; uji McNemar untuk dua sistem.
- Ablasi: segmentasi layout-aware vs paragraf polos; variasi *top-k* (5/10); BM25 vs dense retrieval.

3.4.4 Validasi Ahli & Kegunaan (opsional)

- *Expert review*: 20 sampel jawaban dinilai 3 penilai (benar/parsial/salah) + justifikasi.
- Uji kegunaan ringan (SUS) pada 8–12 pengguna internal untuk prototipe antarmuka tanya–jawab.

3.4.5 Kriteria Penerimaan

- Reader (IndoQA): F1 $\geq$ baseline publik HF.
- End-to-end PDF: F1 $\geq$ 0,60 dan *traceability* $\geq$ 95%.
- Retrieval PDF: Recall@10 $\geq$ 0,85.

3.4.6 Risiko dan Mitigasi

- PDF scan/gambar: dikecualikan; OCR hanya pekerjaan lanjutan.
- Jawaban sangat pendek: atur `max_answer_length` dan normalisasi pasca-proses.
- Batas komputasi: gunakan IndoBERT-Lite, fp16, *gradient accumulation*.

DAFTAR PUSTAKA

[1] Ashish Vaswani et al. “Attention is All You Need”. *Advances in Neural Information Processing Systems*. 2017.

[2] Jacob Devlin et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. *Proceedings of NAACL-HLT* (2019).

[3] Bryan Wilie et al. “IndoBERT: A Pretrained Language Model for Indonesian”. *Proceedings of the 28th International Conference on Computational Linguistics*. 2020.

[4] Patrick Lewis et al. “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks”. *Advances in Neural Information Processing Systems 33*. 2020, pp. 9459–9474.

[5] Jon Saad-Falcon, Daniel Friel, and Krisztian Balog. “PDFTriage: Question Answering over Long, Structured Documents”. *arXiv preprint arXiv:2309.08872*. 2023.

[6] Holy Lovenia, Rahmad Mahendra, et al. “SEACrowd: A Multilingual Multimodal Data Hub and Benchmark Suite for Southeast Asian Languages”. *Proceedings of EMNLP*. 2024.

[7] Bryan Wilie et al. “IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding”. *Proceedings of AACL-IJCNLP*. 2020.

[8] Vladimir Karpukhin et al. “Dense Passage Retrieval for Open-Domain Question Answering”. *Proceedings of EMNLP*. 2020.

[9] Daniel Jurafsky and James H. Martin. *Speech and Language Processing*. 3rd draft. Stanford University, 2025.

[10] Jakarta Research. *IndoQA: Indonesian Question Answering Dataset*. https://huggingface.co/datasets/jakartaresearch/indoqa. 2023.

LAMPIRAN

A Dataset

B Hasil Wawancara

C Rincian Kasus Uji