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, dokumen-dokumen tersebut sangat kompleks karena terdapat banyak bab, istilah teknis, dan rujukan silang sehingga menyulitkan pengguna, karyawan, atau petinggi terkait untuk mendapatkan jawaban yang spesifik secara cepat dan tepat. Melakukan pencarian secara 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 makna dasar tanpa konteks. Pendekatan ini tidak mampu menangkap sinonim, 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]. Tersedia 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 tiga *research gap*: (1) belum ada evaluasi terukur IndoBERT-Lite untuk QA pada dokumen 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. Untuk menjawab gap ini, penelitian memanfaatkan model pralatih `Wikidepia/indobert-lite-squad` sebagai titik awal *fine-tuning* dan dataset `jakartaresearch/indoqa` sebagai sumber untuk 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:

1. 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?
2. 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?
3. 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 kesimpulan.

1.3 Tujuan Penelitian

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

1. 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.
2. Memperoleh model QA IndoBERT-Lite hasil *fine-tuning* pada dataset IndoQA, dan mengukur peningkatan kinerjanya (EM/F1) dibandingkan model pralatih.
3. 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.