BAB I
PENDAHULUAN

1.1 Latar Belakang Masalah

Kegiatan *entry* data dari dokumen cetak seperti formulir, surat resmi, atau dokumen izin distribusi masih menjadi bagian penting dalam pengelolaan informasi di berbagai lembaga. Sayangnya, ketika proses ini dilakukan secara manual, potensi terjadinya kesalahan input data dan keterlambatan pemrosesan menjadi sangat tinggi. Berdasarkan penelitian yang dilakukan oleh Nurhaliza dan Lussiana (2022), dalam konteks verifikasi dokumen pada proses bea cukai alat kesehatan selama pandemi COVID-19, volume dokumen yang tinggi dan keterbatasan tenaga manusia menyebabkan akurasi dan efisiensi menjadi persoalan utama. Untuk mengatasi hal ini, mereka mengimplementasikan metode *Optical Character Recognition* (OCR) untuk mengenali karakter dari dokumen izin distribusi, dan hasilnya menunjukkan tingkat keberhasilan pengenalan mencapai 92%.

Walaupun OCR mampu membaca karakter dari citra dokumen dengan cukup baik, tantangan masih muncul ketika sistem harus mengenali karakter khusus seperti simbol, angka mirip, atau teks yang tercetak tidak jelas. Selain itu, OCR tidak memiliki kemampuan untuk memahami struktur atau makna dari data yang diekstrak. Oleh karena itu, pendekatan berbasis *Natural Language Processing* (NLP) diperlukan untuk menganalisis dan mengidentifikasi konteks dari teks hasil ekstraksi. Akan tetapi, teknik NLP konvensional sering kali belum cukup kuat untuk mengenali urutan data yang kompleks secara semantik. Di sinilah peran metode *Conditional Random Field* (CRF) menjadi relevan, karena CRF merupakan algoritma statistik yang dirancang khusus untuk memproses data berurutan dan memberikan label pada setiap elemen dalam konteks keseluruhan teks.

CRF bekerja dengan mempertimbangkan keterkaitan antar elemen dalam suatu urutan, seperti antar kata atau entitas dalam kalimat, sehingga lebih akurat dalam mengenali dan memvalidasi struktur data. Beberapa studi sebelumnya telah menunjukkan bahwa model CRF mampu meningkatkan performa validasi data, khususnya dalam tugas pengenalan entitas (*Named Entity Recognition*).

Dalam konteks pemrosesan data hasil OCR, CRF dapat digunakan untuk memastikan bahwa urutan entitas seperti nama, tanggal, dan nomor dokumen berada pada posisi dan pola yang sesuai. Dengan integrasi antara OCR sebagai pengenalan teks, NLP sebagai pemroses bahasa, dan CRF sebagai pemvalidasi struktur data, sistem yang dihasilkan akan lebih andal dalam menghindari kesalahan input dan mempercepat proses *entry* data..

Atas dasar urgensi tersebut, penelitian ini bertujuan untuk membangun sistem validasi data yang terotomatisasi dengan memanfaatkan teknologi OCR untuk membaca dokumen, NLP untuk mengidentifikasi struktur teks, dan CRF untuk memverifikasi urutan dan kebenaran entitas dalam data yang diproses. Solusi ini diharapkan dapat menjawab tantangan pada proses entry data manual yang selama ini banyak menyita waktu dan rentan kesalahan.

1.2 Identifikasi Masalah

Dalam laporan tugas akhir ini akan dijabarkan rumusan masalah yang mencakup:

1.2.1 Rumusan Masalah

(1) Bagaimana merancang sistem terintegrasi yang mengombinasikan teknologi Optical Character Recognition (OCR) dan Natural Language Processing (NLP) untuk membaca serta memahami isi dokumen dalam rangka mempermudah proses entri data secara otomatis?

(2) Bagaimana implementasi algoritma Conditional Random Field (CRF) dapat meningkatkan akurasi validasi entitas hasil ekstraksi dokumen dari OCR dan pemrosesan NLP?.

1.2.2 Batasan Masalah

(1) Dokumen yang digunakan dalam pengujian adalah dokumen semi-struktur, berupa invoice , yaitu dokumen transaksi yang umumnya mencakup informasi seperti nomor dokumen , tanggal penerbitan, identitas pelanggan, rincian barang atau jasa, serta total pembayaran

(2) Sistem OCR yang digunakan terbatas pada pengenalan teks berbahasa Indonesia yang dicetak (bukan tulisan tangan).

(3) NLP dalam penelitian ini difokuskan pada proses Named Entity Recognition (NER) untuk mengenali entitas tertentu seperti nama, tanggal, nomor identitas, dan jenis informasi lainnya.

(4) Model CRF yang digunakan adalah supervised learning, di mana data latih telah telah diberi label sebelumnya.

(5) Evaluasi kinerja sistem dilakukan berdasarkan metrik akurasi validasi entitas, tanpa membahas performa komputasi seperti kecepatan eksekusi atau efisiensi memori.

(6) Penelitian ini tidak mencakup pengolahan lanjutan seperti penyimpanan data ke basis data atau integrasi dengan sistem informasi lain.

1.3 Tujuan Penelitian

Sistem ini akan menghasilkan hasil sistem yang mampu mempercepat dan dan meningkatkan akurasi proses entri data dari dokumen fisik. Beberapa tujuan dari sistem yang dikembangkan ini adalah sebagai berikut:

(1) Penyelesaian perancang sistem entri data otomatis dengan integrasi teknologi OCR dan NLP.

(2) Penyelesaian pengimplementasikan algoritma (CRF) untuk mengekstraksi dan memverifikasi entitas dari data hasil pemindaian OCR..

1.4 Manfaat Penelitian

Sistem yang dikembangkan ini dapat membantu mengatasi masalah masalah teknis yang ada, tetapi juga memberikan kontribusi secara teoritis maupun praktis di bidang ilmu komputer, khususnya dalam area pengolahan dokumen, kecerdasan buatan, dan pemrosesan bahasa alami:

1.4.1 Manfaat Teoritis

(1) Pengayaan literatur ilmiah dalam bidang pemrosesan dokumen otomatis, khususnya yang berkaitan dengan digitalisasi dan validasi data berbasis dokumen semi-struktur.

(2) Penerapan metode CRF dalam konteks validasi entitas hasil ekstraksi teks dari dokumen cetak menambah referensi mengenai penggunaan CRF. di luar bidang pengenalan bahasa alami secara umum.

(3) Penerapan gabungan OCR dan NLP sebagai pendekatan terintegrasi yang jarang dibahas dalam satu rangkaian sistem, sehingga dapat menjadi acuan atau fondasi untuk penelitian-penelitian lanjutan di masa depan, terutama dalam sistem cerdas berbasis teks.

(4) Memberikan gambaran teknis dan metodologis tentang bagaimana proses integrasi antar teknologi tersebut dapat dilakukan secara sinergis dan terstruktur.

1.4.2 Manfaat Praktis

(1) Menyediakan solusi sistem validasi data otomatis yang dapat mengurangi beban kerja manual dalam proses entri data dari dokumen fisik, yang selama ini sering menimbulkan kesalahan dan keterlambatan.

(2) Meningkatkan efisiensi operasional di instansi atau organisasi yang mengelola dokumen dalam jumlah besar, seperti lembaga pemerintahan, rumah sakit, institusi pendidikan, maupun perusahaan swasta.

(3) Memberikan alat bantu berbasis teknologi yang dapat membantu operator atau staf administrasi dalam memastikan bahwa data yang dimasukkan ke dalam sistem sudah benar dan tervalidasi sesuai dengan struktur yang diharapkan.

(4) Menjadi dasar penerapan sistem berbasis AI untuk digitalisasi dokumen di lingkungan yang masih belum terdigitalisasi sepenuhnya, khususnya di daerah-daerah atau sektor dengan sumber daya terbatas.

1.5 Kebaruan

Penelitian ini menghadirkan pendekatan baru dalam pengolahan dokumen fisik dengan memadukan tiga teknologi utama: Optical Character Recognition (OCR), Natural Language Processing (NLP), dan metode pembelajaran mesin *Conditional Random Field* (CRF). Meskipun ketiganya telah digunakan secara terpisah dalam berbagai studi, kombinasi ketiganya dalam satu sistem terintegrasi untuk tujuan validasi entri data masih jarang ditemukan dalam konteks penelitian lokal maupun praktis di Indonesia.

1.6 Kerangka Penulisan

Penyusunan TA ini mengikuti pedoman yang telah ditetapkan oleh Fakultas Teknologi dan Desain Universitas Pembangunan Jaya. TA yang disusun terstruktur dalam enam bab yang memadukan kerangka metodologis yang komprehensif dan informatif.

(1) BAB I PENDAHULUAN

Mencakup pemahaman latar belakang masalah, identifikasi masalah yang mendasari penelitian ini, penyusunan rumusan masalah yang terinci, penentuan batasan cakupan masalah, tujuan penelitian yang ingin dicapai, manfaat dari hasil penelitian, serta pembahasan mengenai kebaruan dan kerangka penelitian yang menjadi dasar metodologi.

(2) BAB II TINJAUAN PUSTAKA

Menguraikan konsep dasar teori yang memiliki relevansi signifikan dengan penelitian ini. Bab ini juga merujuk pada penelitian sebelumnya sebagai sumber referensi yang kuat, serta memberikan tinjauan teoritis yang mendalam melalui sub-bab tinjauan pustaka.

(3) BAB III TAHAPAN PELAKSANAAN

Memberikan gambaran menyeluruh mengenai langkah-langkah pelaksanaan penelitian dan metode pengujian.

(4) BAB IV PERANCANGAN

Menguraikan langkah-langkah penelitian dan merinci rancangan pengujian. Metode Analytical Hierarchy Process diimplementasikan dalam pengembangan aplikasi berbasis website, yang menjadi fokus utama penelitian ini.

(5) BAB V HASIL DAN PEMBAHASAN

Memaparkan secara rinci data yang diperoleh dari pengujian menggunakan metode black box dan white box. Analisis hasil pengujian diulas secara mendalam, menyertakan interpretasi dan kesimpulan yang diambil dari temuan-temuan tersebut.

(6) BAB VI PENUTUP

Menampilkan kesimpulan menyeluruh dari seluruh penelitian, mencakup ringkasan temuan, implikasi praktis, dan saran-saran untuk pengembangan lebih lanjut. Bab ini menyajikan kesan akhir dan memberikan arah untuk penelitian masa depan di bidang yang terkait.