

BAB I

PENDAHULUAN

1.1. Latar Belakang

Di era yang semakin berkembang, penggunaan teknologi sudah menjadi bagian dalam kehidupan manusia sehari-hari. Teknologi mampu membantu manusia untuk hidup lebih praktis dan cepat (Novie Tri Lestari et al., 2022). Seiring dengan perkembangan teknologi yang ada, pengelolaan dokumen secara otomatis dan efisien menjadi suatu kebutuhan di berbagai sektor seperti pemerintahan, bisnis, Pendidikan dan berbagai macam sektor lainnya. Hal ini disebabkan karena banyak tantangan yang dihadapi Ketika mengelola dokumen, khususnya dokumen administratif seperti surat perizinan, surat peminjaman, dan surat undangan. Biasanya, dokumen-dokumen ini memerlukan pengelolaan secara manual untuk melakukan klasifikasi dan mengekstrak informasi yang ada dalam dokumen tersebut, seperti jenis surat, informasi kegiatan, tanggal, waktu, Lokasi dan informasi penting lainnya. Pengelolaan secara manual ini merupakan hal yang rentan terhadap kesalahan dan sering mengalami *human error* (Zhang Neng, 2022).

Proses pengelolaan dokumen khususnya di Kapel St Yohanes Rasul, masih dilakukan sepenuhnya secara manual seperti pembacaan dokumen satu per satu, hingga pencatatan informasi. Metode ini tidak hanya memakan waktu, tetapi juga sering kali mengalami kesalahan pencatatan akibat kesalahan pengurus dari Kapel St Yohanes Rasul ataupun kehilangan dokumen karena sebagian besar dokumen yang tersedia merupakan dokumen berbentuk fisik. Sehingga memerlukan teknologi yang dapat mengubah dokumen tersebut menjadi digital dan dapat disimpan di media penyimpanan digital. Salah satu teknologi yang dapat membantu kebutuhan ini adalah *Optical Character Recognition* (OCR), yang dapat membantu melakukan konversi dari dokumen fisik atau tulisan tangan ke dalam format digital (Singh et al., 2024).

Saat ini, sudah terdapat berbagai macam program OCR yang mampu mengkonversi dokumen atau gambar menjadi text digital (Novie Tri Lestari et al., 2022), namun tanpa adanya analisa atau ekstraksi informasi lebih lanjut. Hal ini menyebabkan

pengelolaan dokumen akan semakin sulit apabila di masa depan belum ada teknologi yang juga dapat menganalisa hasil konversi dokumen tersebut. Maka, teknologi seperti *Natural Language Processing* (NLP) dapat membantu proses pengelompokan atau pengklasifikasian dokumen menjadi semakin relevan.

Masalah utamanya adalah kurangnya sistem yang mampu untuk melakukan ekstraksi dokumen, klasifikasi dokumen dan juga mendapatkan informasi yang pada dokumen tersebut secara bersamaan. Hal ini disebabkan karena sistem yang ada hanya mampu menangani aspek aspek tertentu seperti konversi teks menggunakan OCR (Saoji et al., 2021) atau analisa teks menggunakan NLP tanpa adanya integrasi dari kedua hal tersebut. Akibatnya, waktu yang diperlukan untuk pengelolaan dokumen akan memakan waktu yang lama untuk mengetahui informasi yang ada pada dokumen.

Penelitian ini mengusulkan solusi berupa integrasi pada teknologi OCR dan NLP untuk membangun sistem klasifikasi dokumen dan ekstraksi informasi pada dokumen. Sistem ini dirancang untuk melakukan klasifikasi secara otomatis dalam mengkategorikan dokumen menjadi jenis tertentu seperti surat perizinan, surat undangan dan surat peminjaman serta mengekstraksi informasi yang ada pada dokumen tersebut seperti nama kegiatan, lokasi kegiatan, nama yang diundang, daftar barang yang dipinjam dan tanggal peminjaman. Solusi ini diharapkan mampu untuk membantu dalam pengelolaan dokumen sehingga dapat menghemat waktu dan tenaga khususnya pada Kapel St Yohanes Rasul.

1.2. Rumusan Masalah

Berdasarkan latar belakang yang telah disebutkan, terdapat beberapa rumusan masalah yang bisa diidentifikasi sebagai berikut:

1. Bagaimana merancang sistem yang dapat mengintegrasikan teknologi OCR dan NLP untuk melakukan ekstraksi informasi dari dokumen administratif dan mengklasifikasikan jenis dokumen tersebut?
2. Bagaimana tingkat akurasi dan efektivitas sistem OCR dan NLP dalam mengklasifikasikan dan mengekstraksi informasi dari dokumen administratif?

3. Bagaimana cara mengintegrasikan sistem pada platform website yang dapat diakses dengan mudah oleh pengguna, serta memastikan kemudahan dalam penggunaan dan pengoperasian sistem tersebut?

1.3. Tujuan Penelitian

Berdasarkan rumusan masalah yang telah diidentifikasi, maka tujuan dari penelitian ini adalah untuk mengklasifikasikan dokumen dan mengekstraksi informasi informasi yang ada pada dokumen tersebut dengan tujuan spesifik sebagai berikut :

1. Merancang sistem integrasi OCR dan NLP untuk klasifikasi dan ekstraksi informasi dokumen.
2. Mengevaluasi akurasi dan efektivitas sistem OCR dan NLP dalam mengklasifikasikan dan mengekstraksi informasi dari dokumen administratif.
3. Mengintegrasikan sistem yang dikembangkan ke dalam sebuah platform website yang dapat diakses dengan mudah oleh pengguna.

1.4. Batasan dan Asumsi Penelitian

1.4.1. Batasan Penelitian

Penelitian ini memiliki beberapa batasan untuk mengatur ruang lingkup dalam penyelesaian masalah pengelolaan dokumen, diantaranya adalah:

1. Jenis dokumen yang akan diklasifikasi hanya mencakup surat perizinan, surat undangan dan surat peminjaman.
2. Dokumen tersedia dalam format PDF atau citra (JPG/PNG) yang diperoleh dari arsip yang dimiliki Kapel.
3. Sistem yang dikembangkan hanya mendukung dokumen dalam Bahasa Indonesia.

1.4.2. Asumsi Penelitian

1. Kualitas Dokumen

Dokumen yang diunggah diasumsikan memiliki kualitas citra yang memadai sehingga teks di dalamnya dapat dikenali oleh teknologi OCR. Kualitas

memadai ini mencakup resolusi yang cukup, teks yang terbaca jelas, serta tingkat kontras yang baik antara teks dan latar belakang kertas. Menyadari adanya kualitas dokumen yang bervariasi, maka sistem akan dilengkapi dengan tahap pra-pemrosesan sebelum ekstraksi OCR.

2. Struktur Konten Dokumen

Dokumen yang digunakan memiliki format penulisan yang standar, dengan menyajikan informasi yang relatif konsisten atau menggunakan kata kunci yang umum. contohnya, surat undangan menyertakan label eksplisit seperti "Hari", "Tanggal", "Waktu", dan "Tempat" , sementara surat peminjaman memiliki frasa seperti "Daftar Barang yang Dipinjam". Teks pada dokumen juga merupakan hasil cetak (*machine-printed*) dan bukan tulisan tangan. Serta bahasa yang digunakan dalam dokumen adalah Bahasa Indonesia

1.5. Manfaat Penelitian

Berdasarkan tujuan yang hendak dicapai, maka penelitian ini diharapkan mempunyai manfaat dalam pengelolaan dokumen agar lebih baik dan efisien serta diharapkan juga memberikan manfaat baik secara teori maupun praktik di antaranya:

1. Manfaat bagi pengurus Kapel St Yohanes Rasul
 - Mempermudah dalam proses pengelolaan dan ekstraksi dokumen
 - Mengurangi risiko *human error* dalam memahami informasi dari dokumen
2. Manfaat bagi Penelitian
 - Memberikan kontribusi ke bidang penelitian *Document Processing*.
 - Menjadi dasar dalam pengembangan penelitian lebih lanjut seperti klasifikasi dokumen dengan berbagai format atau mendukung dokumen multibahasa.
3. Manfaat bagi Organisasi atau komunitas lainnya
 - Memberikan solusi teknologi yang dapat diadopsi oleh komunitas atau organisasi lain yang memiliki kebutuhan serupa dalam mengelola dokumen seperti gereja, komunitas lokal atau institusi pendidikan.

1.6. Sistematika Penulisan

Untuk menjaga agar penulisan penelitian ini tetap terfokus pada pokok permasalahan dan tidak meluas ke topik lain, serta memudahkan pembaca untuk memahami alur penelitian dan hasil yang diperoleh, maka penulis Menyusun sistematika penulisan sebagai berikut:

BAB I PENDAHULUAN

Dalam bab ini menjelaskan latar belakang penelitian yang mendasari pentingnya klasifikasi dan ekstraksi dokumen administrasi menggunakan OCR dan NLP. Rumusan Masalah, Tujuan Penelitian, Batasan dan Asumsi Penelitian, Manfaat Penelitian dan Sistematika Penulisan juga dijabarkan untuk memberikan gambaran terkait ruang lingkup dan arah penelitian

BAB II LANDASAN TEORI

Dalam bab ini berisi Literatur Terkait Teori, Dasar Teori, dan Alasan Pemilihan Teori

BAB III METODOLOGI PENELITIAN

Pada bab ini dijelaskan secara rinci mengenai metodologi yang digunakan untuk menyelesaikan permasalahan penelitian. Beberapa hal yang akan dibahas dalam bab ini meliputi Sistematika Penyelesaian Masalah, Alat dan Bahan Penelitian, Desain Sistem dan Pengujian dan Evaluasi Sistem

BAB IV PENGUMPULAN DAN PENGOLAHAN DATA

Bab ini merinci proses pengumpulan data penelitian yang berasal dari arsip dokumen Kapel St. Yohanes Rasul. Dijelaskan pula tahapan pengolahan data awal yang mencakup proses ekstraksi teks dari dokumen menggunakan OCR, mulai dari pra-pemrosesan citra, eksekusi OCR, hingga pasca-pemrosesan teks untuk menghasilkan data yang bersih dan siap digunakan untuk pelatihan model.

BAB V ANALISIS DAN PEMBAHASAN

Bab ini merupakan inti dari pembahasan hasil penelitian. Disajikan analisis mendalam dari hasil eksperimen *fine-tuning* model untuk tugas klasifikasi dokumen dan *Named Entity Recognition* (NER). Selain itu, bab ini juga membahas hasil pengujian fungsionalitas sistem secara terintegrasi, termasuk analisis kinerja, efektivitas, dan penerimaan pengguna terhadap aplikasi yang telah dibangun.

BAB VI KESIMPULAN DAN SARAN

Bab terakhir ini berisi kesimpulan yang ditarik dari seluruh hasil analisis dan pembahasan, yang menjawab rumusan masalah penelitian. Disajikan pula beberapa saran yang dapat menjadi acuan untuk pengembangan sistem atau penelitian sejenis di masa mendatang.