5.83%

SIMILARITY OVERALL

SCANNED ON: 21 JUL 2025, 11:38 AM

Similarity report

Your text is highlighted according to the matched content in the results above.

IDENTICAL 2.05%

CHANGED TEXT 3.77%

Report #27591127

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). 1 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 1.2 Identifikasi Masalah Dalam laporan tugas akhir ini akan dijabarkan rumusan masalah yang mencakup: 1.2 1 (1) (2) Rumusan Masalah 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? 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 (2) Sistem OCR yang digunakan terbatas pada pengenalan teks berbahasa Indonesia yang dicetak (bukan tulisan tangan). (3) (4) (5) (6) NLP dalam penelitian ini difokuskan pada proses Named Entity Recognition (NER) untuk mengenali entitas tertentu seperti nama, tanggal, nomor identitas, dan jenis informasi lainnya. Model CRF yang digunakan adalah supervised learning, di mana data latih telah telah diberi label sebelumnya. Evaluasi kinerja sistem dilakukan berdasarkan metrik akurasi validasi entitas, tanpa membahas performa komputasi seperti kecepatan eksekusi atau efisiensi memori. 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) (2) Merancang dan mengembangkan sistem entri data otomatis dengan integrasi teknologi OCR dan NLP. Mengimplementasikan algoritma Conditional Random Field (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. 3 (2) (3) (4) 1.4.2 (1) (2) (3) (4) 1.5 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. 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. Memberikan gambaran teknis dan metodologis tentang bagaimana proses integrasi antar teknologi tersebut dapat dilakukan secara sinergis dan terstruktur. Manfaat Praktis 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. Meningkatkan efisiensi operasional di instansi atau organisasi yang mengelola dokumen dalam jumlah besar, seperti lembaga pemerintahan, rumah sakit, institusi pendidikan, maupun perusahaan swasta. 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. 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. 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 4 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) (3) (4) (5) BAB II TINJAUAN PUSTAKA

Menguraikan konsep dasar teori yang memiliki relevansi signifikan dengan penelitian ini. 1 Bab ini juga merujuk pada penelitian sebelumnya sebagai sumber referensi yang kuat, serta memberikan tinjauan teoritis yang mendalam melalui sub-bab tinjauan pustaka. BAB III TAHAPAN PELAKSANAAN Memberikan gambaran menyeluruh mengenai langkah-langkah pelaksanaan penelitian dan metode pengujian. BAB IV PERANCANGAN Menguraikan langkah-langkah penelitian dan merinci rancangan pengujian. 1 Metode Analytical Hierarchy Process diimplementasikan dalam pengembangan aplikasi berbasis website, yang menjadi fokus utama penelitian ini. 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. 5 (6) BAB VI PENUTUP Menampilkan kesimpulan menyeluruh dari seluruh penelitian, mencakup ringkasan temuan, implikasi praktis, dan saran-saran untuk pengembangan lebih lanjut. 1 Bab ini menyajikan kesan akhir dan memberikan arah untuk penelitian masa depan di bidang yang terkait. 6 2.1 BAB II TINJAUAN PUSTAKA Pencapaian Terdahulu Penelitian ini mendasarkan pengembangannya pada berbagai penelitian terdahulu yang dijadikan sebagai referensi dan rujukan utama. Tabel 2. 1 Pencapaian Terdahulu 7 No. Nama Peneliti Publikasi Hasil 1. Rizal Muhamd Aldi (2022) ImpelmentasiI OCR Dengan Metode Autoencoder Untuk Verifikasi Data KTP. Thesis (S1), FIK/INFO. 22 082. 26. Sep 2022 . 10 Penelitian ini memiliki tujuan agar gambar menjadi lebih bersih dan hasil OCR menjadi lebih akurat. Hal ini di dasari keperluan mendaftar akun dalam webside atau aplikasi 2. Kristian Adi Nugraha. (2024) Penerapan Optical Character Recognition Untuk Pengenalan Variasi Teks Pada Media Persentasi Pembelajaran Jurnal Buana informatika , Volume 15, Nomor 01 ,April 2024 Penelitian ini memiliki latar belakang permasalahan media pembelajaran digital umumnya tersimpan dalam bentuk citra karena memiliki unsur visual bentuk citra dianggap sebagai gambar 3. Firhan Maulana R., Kevin Akbar A, Hendy I. (2021) Indonesia Id Card Extractor Using Optical Character Recognition and Natural Language Post Processing Penelitian ini memiliki

latar belakang untuk meningkatkan akurasi dalam mengekstrak yang menghasilkan tesk dari gambar 2.2 Tinjauan Teoritis Tinjauan teoritis adalah bagian dari penelitian yang menyajikan dan menganalisis kerangka konseptual serta teori-teori yang mendukung atau relevan dengan topik penelitian yang sedang dijalankan. Tujuan dari tinjauan teoritis adalah untuk memberikan dasar pemahaman yang kuat tentang landasan konseptual dan teoritis yang melatar belakangi penelitian. 2.2.1 Optical Character Recognition (OCR) OCR memungkinkan komputer menerjemahkan teks dari gambar atau dokumen cetak menjadi teks digital. Prosesnya melibatkan pra-pemrosesan gambar (seperti binarisasi, perataan, dan penghapusan noise), segmentasi karakter, pengenalan pola, dan pemrosesan ulang hasil ekstraksi Berdasarkan studi yang dilakukan oleh Meharuniza Nazeem dan rekan-rekannya (2024), sejumlah pustaka OCR seperti Tesseract, EasyOCR, MMOCR, dan PaddleOCR terbukti mampu memberikan akurasi tinggi, terutama pada bahasa-bahasa dengan sumber daya terbatas (low-resource languages) 8 4. A Syarif R., Tubagus M A. (2020) Kombinasi Metode NER – OCR untu k Meningkatkan efesiensi Pengambilan informasi Diposter berbahasa indonesia Vol 8 issue 4. Oktober 202o. Penelitian ini memiliki latar Belakang dimana penyelengara Acara di Indonesia sering Menggunakan media poster digital Namun proses manula untuk Menstranfer informasi dari poster Digital ke situs web sering terkenda. 5. Nehru, Yosi Raduas H., Amelinda Callsta D I. (2024) Implementasi cannyfilter Optical Character Recognition (OCR) untuk Identifikasi Tanda Nomor Kendaraan Bermotor Volume 6 No.1 2024 Penelitian ini memiliki latar Belakang memonitoring kendaraan Yang melanggar aturan lalu lintas Tetapi pencatatan masih dilakukan Secara manual oleh operator Tesseract, misalnya, berhasil mencapai tingkat akurasi sebesar 92% untuk teks berbahasa Inggris dan hingga 93% pada beberapa bahasa lokal di India. Di sisi lain, pendekatan terbaru yang menggabungkan arsitektur Transformer secara hibrida, yaitu model DOTA, mampu meningkatkan ketepatan dalam mengenali urutan karakter (sequence OCR) dengan rata-rata akurasi sebesar 77,8% pada dataset uji standar.

Peningkatan ini diperoleh dengan menyisipkan algoritma Conditional Random Field (CRF) sebagai mekanisme pelabelan berbasis urutan (DOTA Consortium, 2025). Kendati demikian, sistem OCR masih menghadapi kendala saat berhadapan dengan dokumen yang buram atau menggunakan jenis huruf yang tidak umum, sehingga dibutuhkan proses koreksi lanjutan untuk memperbaiki hasil ekstraksi teks.. 2.2 4 8 2 Natural Language Processing (NLP) NLP adalah cabang ilmu yang memungkinkan komputer memahami bahasa manusia. Tujuan utama dalam penelitian ini adalah menggunakan NLP untuk memproses teks dari OCR agar bisa dikenali struktur serta konteksnya. Proses NLP yang digunakan meliputi: a. Tokenisasi: memisahkan teks menjadi kata atau frasa, b. POS tagging: memberi label bagian kata berdasarkan fungsi sintaksis, c. Stemming/Lematisasi: menyederhanakan kata ke bentuk dasarnya, d. 4 11 Named Entity Recognition (NER): mengenali dan menandai entitas seperti nama, tanggal, atau nomor. Studi terkini Indigo-BiGRU-CRF (2023) untuk NER teks bahasa Indonesia menunjukkan peningkatan performa, mempertegas bahwa metode sequence modeling seperti CRF sangat efektif untuk dokumen lokal : 2.2.3 Conditional Random Field (CRF) CRF adalah algoritma pembelajaran terawasi yang digunakan untuk memberi label pada urutan data. Pada OCR+ NLP, CRF efektif untuk memprediksi jenis entitas berdasarkan konteks sekeliling token. 9 Menurut Lafferty dkk. (2001), CRF mempertimbangkan ketergantungan label antar kata, sehingga mampu menghasilkan pemetaan entitas yang lebih akurat dibanding model kla-sik. Studi kombinasi BiGRU-CRF pada data teks Bahasa Indonesia (2023) menunjukkan efektivitasnya dalam meningkatkan skor F1 pada tugas ekstraksi informasi dan penelitian klinis CRF juga menegaskan bahwa CRF unggul dalam struktur teks berbasiskan entitas 1. Kelebihan CRF: A. Memperhatikan Konteks secara Menyeluruh CRF mempertimbangkan ketergantungan antar label dalam satu rangkaian data. Hal ini memungkinkan sistem memahami konteks kalimat atau susunan elemen data lebih akurat dibanding model klasifikasi independen seperti Naive Bayes. B. Menghindari Masalah Label yang Tidak Konsisten dengan menggunakan pendekatan global (global normalization), CRF dapat

menghindari kombinasi label yang tidak masuk akal, karena model belajar dari keseluruhan struktur sekuens, bukan dari prediksi token per token saja. C. Fleksibel dalam Penambahan Fitur CRF memungkinkan integrasi berbagai jenis fitur seperti kata sebelumnya, panjang kata, kapitalisasi, dan lainnya tanpa harus membuat asumsi independensi antar fitur. D. Performa Unggul untuk NER dan Validasi Teks Terstruktur berdasarkan studi oleh Lample et al. (2016) dan penelitian lokal seperti Rizka & Harjoko (2021), CRF menunjukkan hasil yang lebih stabil dan akurat dalam menangani teks formal seperti formulir data, artikel, maupun dokumen administratif. 2 Kekurangan CRF: a. Waktu Pelatihan Lebih Lama karena sifatnya yang menghitung dependensi antar label secara menyeluruh, pelatihan model CRF bisa memakan waktu lama, terutama pada dataset besar atau sangat kompleks. 10 b. Memerlukan Ekstraksi Fitur Manual tidak seperti deep learning modern (misalnya LSTM atau BERT), CRF tidak melakukan learning terhadap representasi data. Oleh karena itu, performanya sangat bergantung pada kualitas fitur yang dirancang secara manual oleh peneliti. c. Tidak Cocok untuk Data Tidak Terstruktur atau Ambigu untuk teks yang sangat bebas (seperti komentar media sosial), CRF kurang andal karena struktur datanya sulit dipelajari hanya dari fitur statistik permukaan. d. Kesulitan dalam Generalisasi jika Overfitting bila jumlah fitur terlalu banyak atau tidak relevan, CRF bisa mengalami overfitting terhadap data latih, sehingga kurang efektif pada data uji yang belum dikenali sebelumnya . 2.2.3 Validasi Entri Data Validasi data memastikan bahwa output digital telah sesuai format, struktur, dan isi yang benar sebelum disimpan dan digunakan. Jenis validasi umum mencakup: a) Validasi format (misalnya pola tanggal DD-MM-YYYY), b) Validasi nilai (rentang atau tipe data), c) Validasi logika (kontrol kesesuaian antar entitas, seperti tanggal lahir lebih kecil dari tanggal entri) Dengan integrasi OCR + NLP + CRF, validasi dapat dilakukan secara otomatis den gan mempertimbangkan pola dan konteks, bukan hanya aturan statis. Misalnya, CRF membantu membedakan antara tanggal dan angka seri

berdasarkan urutan token—suatu pendekatan yang lebih adaptif dan kontekstual dibanding validasi manual. 11 2.2.4 Metodologi Pengembangan Sistem Metodologi pengembangan sistem digunakan sebagai panduan dalam merancang, membangun, dan menguji sistem yang diusulkan. Dalam penelitian ini, pendekatan yang digunakan adalah metode prototyping, yaitu metode yang menekankan pada pembuatan versi awal (prototipe) dari sistem yang kemudian diperbaiki secara bertahap berdasarkan masukan pengguna. 1. Pemilihan Metode Prototyping Metode prototyping dipilih karena: 1 Membantu peneliti dan pengguna memperoleh pemahaman awal tentang sistem yang sedang dikembangkan. 2 Memungkinkan evaluasi fungsionalitas sejak tahap awal. 3 Cocok untuk sistem dengan kebutuhan yang dapat berubah atau disesuaikan melalui uji coba langsung. 4 Mendorong partisipasi aktif pengguna dalam siklus pengembangan sistem. Metode ini sangat relevan dalam konteks pengembangan sistem validasi entri data berbasis OCR dan NLP, karena memungkinkan pengujian langsung terhadap fungsionalitas seperti ekstraksi teks, pengenalan entitas, serta mekanisme validasi otomatis sebelum sistem dikembangkan secara penuh. 2. Tahapan Metode Prototyping 1. Pengumpulan Kebutuhan (Requirement Gathering) Tahap awal melibatkan identifikasi kebutuhan sistem berdasarkan studi literatur, observasi proses entry data secara manual, dan konsultasi dengan calon pengguna. 2. Pembuatan Prototipe Awal (Build Prototype) Prototipe sistem awal dikembangkan menggunakan HTML, CSS, dan JavaScript sebagai antarmuka pengguna, serta Python atau Node.js pada sisi server untuk menangani proses OCR, NLP, dan CRF. 12 3. Evaluasi dan Umpan Balik Pengguna (User Evaluation & Feedback) Prototipe diuji oleh pengguna untuk mengamati kinerja sistem, kejelasan tampilan, serta akurasi validasi data. Masukan dari pengguna dicatat untuk perbaikan selanjutnya. 4. Pengumpulan Kebutuhan (Requirement Gathering) Tahap awal melibatkan identifikasi kebutuhan sistem berdasarkan studi literatur, observasi proses entry data secara manual, dan konsultasi dengan calon pengguna. 5. Pembuatan Prototipe Awal (Build Prototype) Prototipe sistem awal dikembangkan

menggunakan HTML, CSS, dan JavaScript sebagai antarmuka pengguna, serta Python atau Node.js pada sisi server untuk menangani proses OCR, NLP, dan CRF. 6. Evaluasi dan Umpan Balik Pengguna (User Evaluation & Feedback) Prototipe diuji oleh pengguna untuk mengamati kinerja sistem, kejelasan tampilan, serta akurasi validasi data. Masukan dari pengguna dicatat untuk perbaikan selanjutnya. 7. Penyempurnaan Sistem (Refinement) Berdasarkan umpan balik, sistem disesuaikan dan diperbaiki hingga mencapai performa dan fungsionalitas yang diharapkan. Proses iteratif ini dapat dilakukan beberapa kali 8. Implementasi Akhir (Final Implementation) Setelah prototipe dianggap stabil dan sesuai dengan kebutuhan, sistem diimplementasikan secara penuh dan diuji secara menyeluruh 13 BAB III TAHAPAN PELAKSANAAN 3.1 Langkah – langkah pelaksanaan Dalam tahapa n pelaksanaan tugas akhir, peneliti menggunakan tabel yang dapat memberikan gambaran yang jelas. Berikut adalah langkah-langkah yang di terapkan: Gambar 3. 1 Langkah Pelaksanaan Pelaksanaan pada penelitian ini dirancang secara terstruktur dengan mengacu pada metode prototyping, yang memungkinkan proses pengembangan sistem dilakukan secara bertahap dan melibatkan interaksi langsung dengan pengguna. Setiap fase dalam proses ini disusun untuk memastikan bahwa solusi yang dibangun mampu mengatasi permasalahan dalam validasi data secara otomatis, dengan memanfaatkan integrasi antara teknologi OCR, NLP, dan metode CRF, Berikut uraian detail setiap tahap: 14 Studi Literatur Analisis Kebutuhan Sistem Perancangan Sistem dan Prototipe Awal Pengembangan Sistem Revisi dan Penyempurnaan Sistem Uji Coba Prototipe dan Evaluasi Awal Implementasi Final dan Dokumentasi Mulai Selesai 1. Studi Literatur dan Observasi Lapangan Tahap awal dilakukan untuk memperoleh pemahaman konseptual dan teknis mengenai teknologi yang akan digunakan. Aktivitas yang dilakukan meliputi: a) Menelaah jurnal-jurnal ilmiah, laporan tugas akhir, dan buku teks yang relevan mengenai OCR, NLP, dan CRF; b) Mengidentifikasi kendala umum dalam proses entri data manual; c) Melakukan observasi di lingkungan nyata (instansi/organisasi) untuk memahami proses entri data

yang saat ini berjalan secara konvensional. 7 2. Analisis Kebutuhan Sistem Tahap ini bertujuan untuk mengidentifikasi kebutuhan pengguna dan merumuskan spesifikasi sistem. Aktivitasnya meliputi: a) Menentukan jenis dokumen yang akan diproses (misalnya formulir isian atau dokumen cetak); b) Menentukan data yang perlu diekstrak dan divalidasi (misalnya: nama, NIK, tanggal, kode tertentu); c) Menyusun use-case diagram atau skenario penggunaan sistem; d) Merancang kebutuhan fungsional (apa yang harus dilakukan sistem) dan non- fungsional (kecepatan, akurasi, kemudahan penggunaan). 3. Perancangan Sistem dan Prototipe Awal Pada tahap ini dilakukan perancangan awal sistem, baik dari sisi antarmuka maupun alur data. Rancangan mencakup: a) Arsitektur sistem berbasis client-server; b) Desain tampilan antarmuka pengguna menggunakan HTML/CSS/JS; c) Alur kerja mulai dari input gambar, proses OCR, pengolahan teks NLP, hingga validasi data dengan CRF; d) Diagram alir proses (flowchart), dan diagram konteks sistem. 4. Pengembangan Sistem Sistem dibangun berdasarkan desain yang telah disepakati. Pengembangan dilakukan secara modular agar tiap komponen bisa diuji secara terpisah. Beberapa tahapan penting meliputi: 15 a) Implementasi proses OCR menggunakan Tesseract untuk ekstraksi teks dari gambar; b) Penerapan NLP (tokenisasi, normalisasi, POS tagging) menggunakan pustaka Python seperti spaCy atau NLTK; c) Integrasi algoritma CRF untuk mengenali pola entitas dan memvalidasi data; d) Pembuatan antarmuka web interaktif menggunakan HTML, CSS, dan JavaScript agar pengguna dapat mengunggah gambar dan melihat hasil validasi secara langsung. 5. Uji Coba Prototipe dan Evaluasi Awal Prototipe diuji dengan dataset awal berupa dokumen gambar yang telah dipilih sebelumnya. Tujuannya adalah: a) Memastikan alur kerja sistem sudah berjalan sesuai rencana; b) Menilai keakuratan hasil ekstraksi dan validasi; c) Mencatat bug atau kelemahan sistem yang perlu diperbaiki; d) Mengumpulkan masukan dari pengguna tentang antarmuka dan kemudahan penggunaan. 6. Revisi dan Penyempurnaan Sistem Setelah mendapat umpan balik dari pengguna dan hasil pengujian awal, dilakukan perbaikan pada

bagian-bagian sistem yang kurang optimal. Langkah ini bisa berlangsung dalam beberapa iterasi tergantung kompleksitas masalah, dan mencakup: a) Perbaikan pada proses ekstraksi teks yang kurang akurat; b) Penyesuaian model CRF agar lebih tepat dalam mengenali pola data; c) Penyempurnaan tampilan atau responsivitas sistem. 7. Implementasi Final dan Dokumentasi Setelah sistem dinyatakan stabil dan memenuhi kebutuhan pengguna, dilakukan: 16 a) Pengujian akhir pada sistem secara keseluruhan; b) Penyusunan dokumentasi teknis dan laporan penelitian; c) Penyimpulan terhadap performa sistem berdasarkan hasil uji dan pengamatan. . 3.2 Metode Pengujian Dalam menguji kualitas perangkat lunak peneliti mengunakan metode pengujian: 3.2.1 Balck box testing Metode pengujian black box berfokus pada pengujian fungsi aplikasi tanpa memerlukan pemahaman tentang struktur kode internal. 9 Tujuannya adalah untuk menjamin bahwa aplikasi beroperasi sesuai dengan spesifikasi yang telah ditentukan. Metode ini menguji semua input dan menentukan apakah output sesuai dengan harapan. Berikut adalah proses pengujian black box testing: Tabel 3. 1 Pengujian Black box testing NO Fitur Keterangan Hasil Sistem 1 Upload dokumen gambar Pengguna memilih file gambar dari perangkat lokal dan menekan tombol "Upload" berformat JPG/PNG Gambar berhasil diunggah dan ditampilkan 2 Ekstraksi teks (OCR) Pengguna klik tombol “Proses OCR” setelah gamba r tampil Sistem menampilkan teks hasil ekstraksi dari gambar 3 Proses NLP dan normalisasi data Pengguna klik tombol “Proses NLP” untu k membersihkan dan menstrukturkan teks Sistem menampilkan teks yang telah dibersihkan dan dipisahkan per entitas 4 Validasi data menggunakan CRF Pengguna klik tombol “Validasi Data” Sistem menandai dan menampilkan dat a yang valid/tidak valid dengan label CRF 5 Export hasil validasi ke Excel/PDF Tombol “Export” diklik File hasil export berhasil diundu h 17 3.2.2 White box testing White box testing melakukan pengujian struktur internal beserta kode program. Penguji melakukan pengujian dengan pemahaman mendalam tentang kode sumber dan logika program. Tujuan utama dari white box testing adalah untuk memeriksa alur kontrol, jalur data,

kondisi, dan logika program secara keseluruhan. Berikut adalah tabel untuk proses pengujian white box testing pada implementasi algoritma CRF Tabel 3. 2 White Box Testing No Algoritma 1 Pengujian Modul OCR (Tesseract API) 2 Pengujian Modul NLP (Tokenisasi, Stopword) 3 Pengujian Modul CRF Validation 3.2.3 Pemecahan Masalah dan Strategi Implementasi Dalam penelitian ini, pendekatan pemecahan masalah dilakukan dengan merancang sistem terintegrasi yang menggabungkan Optical Character Recognition (OCR), Natural Language Processing (NLP), dan algoritma Conditional Random Field (CRF) secara komprehensif dalam satu kerangka kerja. Kebaruan dari pendekatan ini terletak pada integrasi langsung antara ketiga komponen tersebut, yang belum banyak diterapkan secara lengkap dalam konteks validasi data invoice di Indonesia. Sebagian besar studi sebelumnya hanya mengadopsi dua komponen secara terpisah, seperti Rusli et al. (2020) yang menggunakan OCR dan NLP secara sederhana tanpa penguatan model berbasis CRF, serta penelitian oleh Permana (2019) dan Husein (2022) yang mengimplementasikan CRF namun tanpa mengintegrasikan proses akuisisi data melalui OCR dan pemrosesan semantik dari NLP 1. Kategori Entitas yang Dievaluasi: Untuk setiap invoice, sistem diminta mengenali entitas berikut: a. Tanggal Invoice b. Nomor Invoice c. Tanggal Terima d. Deskripsi Barang e. Jumlah/QTY f. Harga g. Total Invoice 2 Label Manual No Nama Entitas Nilai yang Benar 1 Invoice Date 02.07.2025 2 Invoice Number BA10007527 3 Received Date (tidak tersedia) 4 Description Item 1 HF10TQ-P IPP FILM - PRIME 5 Description Item 2 HY3.8FY-P YARN 3.8 - PRIME 6 Description Item 3 BI5.0GA-P ICP INJECTION 5 - PRIME 7 Qty Item 1 15,000 8 Qty Item 2 55,000 9 Qty Item 3 20,000 10 Price Item 1 17,051 11 Price Item 2 15,829 12 Price Item 3 18,198 13 Total Amount 1,654,255,200 a. Hasil Pengujian Sistem NER: Entitas yang seharusnya dikenali (Ground Truth): 13 Entitas yang berhasil dikenali dengan benar (True Positive): 12 Entitas yang dikenali tapi salah (False Positive): 1 18 Entitas yang terlewat / tidak dikenali (False Negative): 0 b. Rumus Evaluasi

: Precision = TP / (TP + FP) = 12 / (12 + 0) = 100% Recall = TP / (TP + FN) = 12 / (12 + 1) = 92.3% F1-Score = 2 × (Precision × Recall) / (Precision + Recall) = 2 × (1.0 × 0.923) / (0.923 + 0.923) ≈ 96% Accuracy = (Jumlah entitas benar / Total entitas yang diuji) = 12 / 13 = 92.3% Dari hasil di atas, error sebesar 8% terjadi bukan karena salah ekstraksi nilai, melainkan karena entitas “Tangga l Terima” tidak tersedia dalam teks (kosong pada dokumen PDF atau tidak terbaca oleh OCR). Penyebab utama ketidakakuratan: Entitas tidak dicetak dalam dokumen (field kosong) → sistem tidak bisa menebak tan pa informasi eksplisit. Variasi posisi dan label field di tiap invoice → sistem perlu pendekatan lebih fleksibel terhadap layout. OCR gaga l membaca teks di area dengan resolusi rendah atau watermark/stempel manual. 19 BAB IV PERANCANGAN 4.1 Analisis Sistem terdahulu Sebelum merancang sistem validasi entry data berbasis integrasi OCR, NLP, dan metode CRF, dilakukan terlebih dahulu kajian terhadap beberapa sistem yang telah dikembangkan pada penelitian sebelumnya. Tujuannya adalah untuk mengidentifikasi kelemahan dan keterbatasan pendekatan terdahulu serta sebagai dasar dalam merancang solusi yang lebih komprehensif dan efektif. Salah satu sistem yang relevan dikembangkan oleh Nurhaliza dan Lussiana (2022), di mana mereka membangun sistem pengenalan karakter otomatis menggunakan teknologi OCR untuk memproses dokumen izin distribusi alat kesehatan. Dengan memanfaatkan pustaka Tesseract OCR, sistem tersebut berhasil memperoleh tingkat akurasi sebesar 98,78% dalam proses pembacaan karakter. Meskipun begitu, sistem tersebut hanya sebatas membaca karakter dan belum dilengkapi dengan mekanisme validasi struktur data maupun pemahaman konteks, sehingga masih memiliki keterbatasan ketika dihadapkan pada dokumen dengan kualitas gambar yang rendah, adanya noise, atau penggunaan jenis huruf yang tidak umum. Sistem lain yang dikembangkan oleh Rusli et al. (2020) menggunakan kombinasi antara OCR dan pendekatan NLP sederhana untuk mengekstrak informasi dari dokumen KTP. Sistem ini mampu melakukan segmentasi data dan normalisasi teks secara

lebih baik, namun proses validasinya masih mengandalkan aturan tetap (rule-based) yang sulit beradaptasi dengan variasi struktur dokumen. Selain itu, belum terdapat kemampuan untuk memverifikasi hubungan antar entitas dalam teks, seperti urutan logis antara NIK, nama, dan alamat. Seiring berkembangnya teknologi pemrosesan bahasa alami, pendekatan menggunakan Conditional Random Field (CRF) mulai banyak diterapkan untuk tugas pengenalan entitas (NER). Model CRF terbukti efektif dalam mengolah teks berurutan dan memberikan label berdasarkan konteks 20 kata sebelumnya maupun sesudahnya. Dalam beberapa penelitian, seperti yang dilakukan oleh Permana (2019) dan Husein (2022), CRF mampu meningkatkan akurasi pelabelan entitas dalam dokumen berbahasa Indonesia. Namun, integrasi metode CRF dalam sistem yang menggabungkan hasil ekstraksi OCR dan pemrosesan NLP masih jarang dikembangkan secara menyeluruh. Dari berbagai studi tersebut, dapat disimpulkan bahwa sebagian besar sistem terdahulu masih fokus pada satu aspek teknologi, baik OCR maupun NLP, dan belum banyak yang menggabungkan seluruh elemen—OCR, NLP, dan CRF—dalam satu sistem yang utuh. Maka dari itu, penelitian ini bertujuan mengembangkan sistem validasi entry data secara otomatis dan menyeluruh dengan menggabungkan ketiga pendekatan tersebut, guna meningkatkan akurasi, mengurangi potensi kesalahan input, dan mempercepat proses validasi data dari dokumen semi-struktur seperti invoice.Berikut adalah tabel untuk proses pengujian white box testing pada implementasi algoritma CRF 4.2 Spesifikasi Kebutuhan Sistem Baru Agar sistem validasi entry data dapat berjalan dengan optimal, maka diperlukan penjabaran kebutuhan sistem yang akan dikembangkan. Bagian ini menyajikan uraian mendetail mengenai spesifikasi teknis dan fungsional dari sistem baru yang dirancang untuk menggantikan proses manual dalam mengekstraksi informasi dari dokumen invoice. 5 Spesifikasi ini mencakup kebutuhan proses, struktur data, pengguna, serta spesifikasi perangkat keras dan lunak yang dibutuhkan untuk menjalankan sistem secara efisien. 4.2.1 Spesifikasi Proses Spesifikasi proses merinci alur operasional sistem yang dikembangkan, mulai dari input

pengguna hingga keluaran hasil akhir. Proses dalam sistem validasi entry data berbasis OCR, NLP, dan metode CRF ini terdiri atas beberapa tahapan penting yang saling terhubung secara sistematis untuk memastikan ketepatan ekstraksi informasi dari dokumen invoice. Berikut ini adalah penjabaran dari tahapan proses yang dilakukan: 21 1. Pengunggahan Dokumen Invoice Pengguna (admin/operator) melakukan upload dokumen invoice hasil pemindaian, baik dalam format gambar (.jpg, .jpeg, .png) maupun PDF. Sistem memverifikasi jenis file agar sesuai dengan format yang didukung. Jika file berupa gambar, maka akan dilanjutkan ke proses OCR. 2. Pemrosesan OCR (Optical Character Recognition) Setelah file berhasil diunggah, sistem akan mengekstraksi teks dari gambar menggunakan Tesseract.js. Tahapan ini mengubah konten visual dokumen menjadi data teks mentah yang dapat dianalisis lebih lanjut. 3. Pembersihan dan Analisis Teks (NLP Preprocessing) Teks hasil OCR selanjutnya diproses melalui tahapan NLP, seperti normalisasi teks, penghapusan karakter yang tidak relevan, dan segmentasi baris. Tujuan dari tahap ini adalah menyederhanakan struktur teks agar lebih siap untuk diekstraksi menggunakan algoritma CRF. 4. Ekstraksi Data menggunakan CRF (Conditional Random Field) Algoritma CRF digunakan untuk mengenali pola dan entitas penting dari teks, seperti: tanggal invoice, nomor invoice, deskripsi barang, jumlah, harga satuan, total, hingga nilai total keseluruhan. Algoritma ini bekerja dengan mengenali hubungan antar kata dan konteksnya berdasarkan pelatihan data sebelumnya. 5. Menampilkan dan Menyimpan Hasil Ekstraksi Data yang berhasil diekstraksi secara otomatis akan ditampilkan dalam form auto-fill di halaman web. Pengguna dapat melakukan pengecekan ulang dan mengklik tombol "Save" untuk menyimpan data ke dalam tabel hasil validasi. 22 6. Tabel Ringkasan dan Pengelolaan Data Setelah data disimpan, sistem akan menampilkan semua hasil OCR & NLP dalam tabel rekap yang dapat dikelola oleh pengguna (misalnya menghapus baris tertentu jika ada kesalahan). 4.2.2 Spesifikasi Data Dalam merancang sistem validasi entry data menggunakan kombinasi teknologi OCR, NLP, dan

algoritma Conditional Random Field (CRF), sangat penting untuk mendefinisikan spesifikasi data yang digunakan. Subbab ini menguraikan jenis, struktur, dan sumber data yang menjadi dasar pengolahan dalam sistem yang dikembangkan.. 1) Jenis Data yang Digunakan Sistem ini dirancang untuk memproses data dalam bentuk dokumen invoice hasil pemindaian (scan) yang biasanya berformat JPG, PNG, atau PDF. Dokumen tersebut mengandung informasi tidak terstruktur yang perlu diubah menjadi bentuk yang dapat dianalisis dan disimpan dalam sistem. Data utama yang diekstraksi meliputi: a) Tanggal Invoice b) Nomor Invoice c) Tanggal Terima d) Deskripsi Barang e) Jumlah (QTY) f) Harga per Unit g) Total Invoice e) Total Amount Setelah dilakukan pemrosesan OCR, data yang awalnya berupa gambar akan dikonversi menjadi teks mentah yang akan diolah lebih lanjut menggunakan NLP dan metode CRF... 23 2) Sumber dan Teknik Pengumpulan Data Untuk membangun model CRF yang akurat, sistem membutuhkan dataset pelatihan yang telah diberi anotasi (labeled data). Dalam tugas akhir ini, pengumpulan data dilakukan dari: a) Dataset invoice hasil scan internal b) Dokumen invoice sample dari berbagai perusahaan c) Data manual yang ditulis ulang dan dianotasi sesuai kebutuhan pelabelan CRF Setiap baris data dianotasi dengan label seperti DATE, INVOICE_NO, ITEM_DESC, QTY, PRICE, TOTAL, dan AMOUNT, agar CRF dapat mengenali pola hubungan antar kata.. 3) Format dan Struktur Dataset Data pelatihan untuk CRF biasanya diformat dalam bentuk baris teks, di mana setiap kata memiliki label-nya masing-masing. Format umum yang digunakan: Tabel 4. 1 Format Umum Styrene ITEM_DESC Monomer ITEM_DESC 1000 QTY 15000 PRICE 15000000 TOTAL Setiap baris data dianotasi dengan label seperti DATE, INVOICE_NO, ITEM_DESC, QTY, PRICE, TOTAL, dan AMOUNT, agar CRF dapat mengenali pola hubungan antar kata.. 4) Perhitungan dan Pelatihan algoritma CRF Dalam metode CRF, proses pelatihan dilakukan menggunakan pendekatan probabilistik. CRF menghitung kemungkinan hubungan antar kata berdasarkan fitur yang dikandung masing-masing token, misalnya: 24 a) Bentuk kata (huruf besar semua,

angka, gabungan huruf dan angka) b) Posisi dalam baris c) Kata sebelum dan sesudahnya CRF akan mencari parameter θ (theta) yang memaksimalkan kemungkinan prediksi label Y terhadap input X. Secara matematis, CRF memaksimalkan: $P(Y|X) = 1 Z(X) \exp(\sum$ $\sum \square k kif k (y i - 1, y i, \square, i))$ Di mana: a) adalah fungsi fitur b) adalah bobot parameter c) Z(X) adalah fungsi normalisasi untuk semua kemungkinan label Pelatihan dilakukan menggunakan library Python seperti sklearn-crfsuite atau PyCRFSuite, lalu hasil model digunakan untuk mengekstrak data dari hasil OCR secara otomatis. 4.2.3 Spesifikasi Data Sistem ini dirancang untuk digunakan oleh satu jenis aktor, yaitu Administrator/Operator: Bertugas mengunggah dokumen invoice, memverifikasi hasil ekstraksi data, dan menyimpan informasi ke sistem. Pengguna diharapkan memiliki pengetahuan dasar tentang pengelolaan dokumen digital serta pemahaman sederhana terhadap elemen-elemen invoice.Sistem dirancang agar mudah digunakan bahkan oleh pengguna dengan latar belakang non-teknis. Tabel 4. 2 Spesifikasi Data 25 No User Penjelasan 1 Admin Admin akan memiliki hak akses untuk mengelola data yang sudah di upload. 2 Operator Operator akan memiliki akses yang terbatas dibanding dengan admim hanya bisa meng upload invoice dan menyimpan datanya. 4.2.4 Spesifikasi Perangkat Keras Agar sistem dapat berjalan lancar, dibutuhkan perangkat keras dengan spesifikasi minimum sebagai berikut: a) Prosesor: Intel Core i3 atau setara b) RAM: Minimal 4 GB c) Penyimpanan: Tersedia ruang kosong minimal 500 MB d) Display: Resolusi minimal 1366x768 pixel e) Koneksi Internet: Dibutuhkan untuk akses pustaka eksternal (seperti Tesseract.js dan model NLP) Spesifikasi tersebut cukup untuk mendukung aktivitas upload file, pemrosesan teks, dan tampilan antarmuka berbasis web. 4.2.5 Spesifikasi Perangkat Lunak Komponen perangkat lunak yang dibutuhkan dalam pengembangan dan implementasi sistem ini mencakup: 26 No Tampilan Fitur Aktor 1 Halaman Login Menampilkan form email dan password untuk mengakses halaman dashboard. Admin, Operator 2 Halaman Dashboard Menampilkan fitur yang

digunakan dalam sistem dan menampilkan tampilan selamat datang. Admin, Operator 3 Halaman validasi Berisi upload invoice dan merubah gambar menjadi text,. Admin, manager Tabel 4. 3 Spesifikasi Perangkat Lunak Sistem Operasi Windows 10 / Linux Ubuntu / macOS Web Browser Google Chr ome Bahasa Pemrograman HTML, CSS, JavaScript Text Editor Visual Studio Code atau setara Library Tambahan Bootstrap 5, Tesseract.js, dan CRF model yang di-load via JavaScript atau Python 4.2.6 Spesifikasi Kebutuhan Sistem Baru Pada tahap ini, dilakukan penyusunan rancangan sistem yang bertujuan untuk mengembangkan aplikasi validasi entry data secara otomatis. Rancangan ini menjadi pondasi awal sebelum sistem dikembangkan dan diimplementasikan secara menyeluruh. Perancangan dilakukan berdasarkan spesifikasi kebutuhan yang telah dianalisis sebelumnya, dengan fokus pada integrasi teknologi OCR (Optical Character Recognition), pemrosesan NLP (Natural Language Processing), serta penerapan algoritma Conditional Random Field (CRF) untuk akurasi ekstraksi data yang lebih tinggi. Subbab ini memuat rancangan visual serta logika dari alur sistem yang akan dibangun, mulai dari flowchart proses hingga tampilan antarmuka aplikasi.. 4.2.7 Spesifikasi Kebutuhan Sistem Baru Pada tahap ini, dilakukan penyusunan rancangan sistem yang bertujuan untuk mengembangkan aplikasi validasi entry data secara otomatis. Rancangan ini menjadi pondasi awal sebelum sistem dikembangkan dan diimplementasikan secara menyeluruh. Perancangan dilakukan berdasarkan spesifikasi kebutuhan yang telah dianalisis sebelumnya, dengan fokus pada integrasi teknologi OCR (Optical Character Recognition), pemrosesan NLP (Natural Language Processing), serta penerapan algoritma Conditional Random Field (CRF) untuk akurasi ekstraksi data yang lebih tinggi. Subbab ini memuat rancangan visual serta logika dari alur sistem yang akan dibangun, mulai dari flowchart proses hingga tampilan antarmuka aplikasi.. 27 4.3.1 Flowchart Metode CRF Flowchart menggambarkan langkah-langkah alur kerja sistem secara urut dari awal hingga akhir. Proses dimulai dari pengunggahan gambar invoice oleh pengguna, dilanjutkan dengan pengenalan teks menggunakan OCR, lalu

pemrosesan teks menggunakan NLP untuk normalisasi dan tokenisasi. Gambar 4. 1 Flowchart Metode CRF 28 Upload Gambar Invoice OCR Ekstraksi Teks Preprocessing NLP Pemodelan Fitur untuk CRF CRF Ekstraksi Entity Validasi dan Tampilkan ke Form Data Valid? Simpan ke Tabel Ya Edit Manual Tidak Mulai Selesai (1) Sistem diaktifkan, dan pengguna mengakses halaman validasi invoice. (2) Pengguna mengunggah file invoice hasil pemindaian dengan format gambar (JPG, PNG) atau PDF. (3) Sistem menjalankan proses Optical Character Recognition (OCR) untuk mengubah gambar menjadi teks digital. (4) Hasil teks dari OCR dibersihkan dari noise, kesalahan spasi, karakter tak dikenal, dan dilakukan tokenisasi (pemecahan kata). (5) Data teks diproses dengan teknik NLP seperti Named Entity Recognition (NER), POS Tagging, dan pattern matching untuk menyiapkan input ke CRF. (6) Setiap token dianalisis dan dilabeli berdasarkan model CRF untuk menentukan entitas seperti: tanggal, nomor invoice, deskripsi barang, jumlah, harga, total, PPN, dan PPH. (7) Hasil label CRF diverifikasi oleh sistem (dan/atau pengguna) untuk mengecek akurasi output. (8) Sistem secara otomatis mengisi field isian dengan hasil ekstraksi CRF ke dalam antarmuka pengguna (form). (9) Jika validasi selesai, pengguna dapat menyimpan data ke tabel dan menyimpan ke dalam database. (10) Proses validasi dan ekstraksi selesai. Pengguna dapat memproses invoice berikutnya. 29 4.3.2 Use Case Flowchart menggambarkan langkah-langkah alur kerja sistem secara urut dari awal hingga akhir. Proses dimulai dari pengunggahan gambar invoice oleh pengguna, dilanjutkan dengan pengenalan teks menggunakan OCR, lalu pemrosesan teks menggunakan NLP untuk normalisasi dan tokenisasi. Gambar 4. 2 Use Case Diagram proses OCR + NPL metode CRF Agar informas i tentang skenario Use Case dapat disajikan dengan jelas dan mudah dipahami, maka perlu dibuat tabel skenario. Tabel skenario harus dibuat dengan pihak yang terlibat, nama skenario, ringkasan singkat, skenario normal, dan skenario alternatif. Tabel skenario ini harus dibuat agar informasi tentang skenario use case jelas dan mudah dipahami. 30 Tabel

4. 4 Login Tabel 4. 5 Menu Dasboar Tabel 4. 6 Menu Validasi 31
Use Case Halaman Login Penjelasan Admin dan operator dihadapkan dengan halaman login. Skenario Utama 1. 2. Admin dan operator masuk ke halaman login. Form login berhasil ditampilkan oleh sistem. Skenario Alternatif 1. 2. User mengalami kesalahan saat memasukkan email/password. Sistem menampilkan pesan error. Kondisi Akir User berhasil masuk kedalam aplikasi setelah melakukan validasi login. Use Case Halaman Utama/ Dashboard Penjelasan Admin dan operator mengakses halaman utama/dashboard aplikasi. Skenario Utama 1. 2. Aktor membuka aplikasi. Sistem menampilkan halaman utama/dashboard dengan informasi dan navigasi ke fitur-fitur utama. Skenario Alternatif 1. 2. Aktor mengalami kesalahan saat membuka aplikasi. Sistem menampilkan pesan error. Kondisi Akir Aktor berada di halaman utama aplikasi Use Case Halaman Validasi Penjelasan Admin dan operator mengakses validasi aplikasi. Skenario Utama 1. 2. 3. 4. Admin dan operator memilih menu “Validasi data”. Admin dan oper tor memilih menu “Data . Admin dan operator memilih menu “Validasi Invoice sesuai PT Akto r memiliki akses untuk mengubah, menambahkan dan . Save data Skenario Alternatif 1. 2. Aktor mengalami kesalahan saat membuka aplikasi. Sistem menampilkan pesan error. Kondisi Akir Aktor berada di halaman validasi data Tabel 4.7 Menu Upload Data Tabel 4. 8 Menu Proses Validasi
4.3.3 Use Case Setelah perancangan Use Case diselesaikan, langkah berikutnya yang dilakukan oleh peneliti adalah menyusun activity diagram untuk masing-masing aktivitas yang terdapat di dalamnya. Diagram ini berfungsi untuk memvisualisasikan alur proses kerja sistem dalam setiap aktivitas yang berlangsung. Selanjutnya, peneliti menguraikan activity diagram berdasarkan peran atau pengguna yang terlibat dalam sistem, guna memberikan gambaran yang lebih rinci mengenai interaksi antar komponen dalam proses bisnis tersebut. 32 Use Case Mengelola Upload data Penjelasan Admin dan operator mengakses upload data. Skenario Utama 1. 2. 3. Admin dan operator memilih enu “Upload data”. Aktor memiliki akses untuk upload data Gambar ditampilkan di table preview, Skenario

Alternatif 1. 2. Aktor mengalami kesalahan saat meng upload. Sistem menampilkan pesan error. Kondisi Akir Aktor berada di halaman validasi data Use Case Mengelola OCR + NLP metode CRF Penjelasan Admin d an operator mengakses extrak dari gambar ke text. Skenario Utama 1. 2. 3. Admin dan operator memilih menu “Scan”. Proses extrak berjalan Hasi text akan masuk sesuai dengan tabelnya Skenario Alternatif 1. 2. Aktor mengalami kesalahan saat extrak. Sistem menampilkan blank table. Kondisi Akir Aktor berada di halaman validasi data 1. Diagram Activity Login Gambar 4. 3 Diagram proses login 2. Diagram Aktivity Validasi Data Menu Diagram aktivity ini menjelaskan urutan proses ketika operator memilih menu mulai dashboard ke tampilan menu validasi data. Diagram aktivity dashboard operator digambarkan pada Gambar berikut Gambar 4. 4 Diagram proses menu validasi 33 3. Diagram Activity Validasi Data OCR + NLP metode CRF Diagram aktivity ini menunjukkan urutan prosedur ya ng digunakan ketika operator meng upload data. Data dapat ditambahkan operator. Aktivitas upload data ini akan mengahilakan data berupa text yang akan di simpan atau di edit sesuai table yang sudah di sediakan digambarkan pada gambar 4.16 berikut Gambar 4. 5 Diagram proses validasi invoice 2 34 4.3.4 Sequence Diagram Sequence diagram merupakan salah satu jenis diagram dalam Unified Modeling Language (UML) yang berfungsi untuk menggambarkan interaksi antar objek dalam sebuah sistem berdasarkan urutan ejadian waktu. Diagram ini memperlihatkan komunikasi antar objek melalui pengiriman pesan selama proses pelaksanaan suatu skenario atau use case. Setiap objek direpresentasikan oleh garis vertikal yang dikenal sebagai lifeline, sementara pesan yang dikirimkan antar objek digambarkan melalui panah horizontal yang menghubungkan antar lifeline. Sequence diagram memudahkan pemahaman terhadap alur proses sistem, menetapkan peran serta tanggung jawab masing-masing objek, dan menyajikan dokumentasi skenario dinamis secara sistematis dan terstruktur. 1. Sequence Diagram Validasi Data Gambar 4. 6 Sequence Diagram Validasi Data 35 BAB V HASIL DAN PEMBAHASAN Pembahasan ini akan menyertakan hasil dari penelitian terkait

validasi data dengan menggunakan OCR + NLP dan metode CRF dengan rinci . 5.1 Hasil Setelah dilakukan implementasi sistem validasi entry data menggunakan integrasi OCR, NLP, dan algoritma CRF, diperoleh hasil yang menunjukkan bahwa sistem mampu mengenali dan memproses dokumen invoice dengan baik. Sistem berhasil mengekstrak informasi penting seperti tanggal invoice, nomor invoice, tanggal diterima, deskripsi produk, jumlah (QTY), harga, total harga, dan total keseluruhan (Total Amount) dari hasil scan gambar invoice yang berformat JPEG atau PNG. OCR berperan dalam mengonversi gambar menjadi teks mentah, yang kemudian diolah lebih lanjut oleh Natural Language Processing (NLP) untuk mengidentifikasi struktur kalimat atau frasa penting. Tahap akhir, Conditional Random Field (CRF) digunakan untuk menandai dan mengklasifikasikan bagian-bagian spesifik dari teks (seperti label INVOICE_DATE, PRICE, QTY, dan TOTAL) berdasarkan urutan kata serta konteksnya. Keberhasilan sistem dapat dilihat dari akurasi data yang ditampilkan secara otomatis pada form input serta keberhasilannya dalam menyimpan dan menampilkan data tersebut ke dalam tabel histori. 5.1.1 Hasil Perancangan Tampilan Desain tampilan website sistem dirancang dengan memperhatikan prinsip kemudahan penggunaan (UI) dan pengalaman pengguna (UX). Dengan demikian, antarmuka website diharapkan menjadi lebih intuitif dan user-friendly. Gambar 5. 1 Menu Login 36 Menu login adalah antarmuka awal dimana admin dan manager masuk ke sistem dan memastikan keamanan akses. Halaman ini memastikan identitas pengguna dan memberikan akses yang sesuai ke sistem. Setelah mengisi informasi login yang benar, pengguna dapat masuk ke sistem dan mengakses fitur-fitur tertentu. Gambar 5. 2 Tampilan dashboard Tampilan utama yang berhasil dirancang dan diimplementasikan adalah halaman Dashboard. Pada bagian ini, pengguna dapat melihat ringkasan dan tabel hasil validasi data invoice yang telah diproses melalui OCR, NLP, dan metode CRF Gambar 5. 3 Tampilan Valiadasi 37 Tampilan Halaman Validasi Invoice CAP yang ditampilkan pada halaman validasi invoice CAP merupakan bagian dari implementasi sistem berbasis web yang dirancang untuk

mendukung proses verifikasi data invoice secara otomatis. Sistem ini memanfaatkan teknologi OCR dan NLP untuk membaca data dari hasil scan dokumen invoice, yang kemudian ditampilkan ke dalam form isian secara otomatis. Bagian atas halam n memuat judul "Vali asi Invoice CAP" yang menunjukkan konteks spesifik halaman ini, yaitu untuk menangani validasi invoice milik Chandra Asri Petrochemical (CAP). Selanjutnya, pengguna disuguhkan dengan fitur unggah dokumen, yang memungkinkan mereka untuk memilih lebih dari satu file dengan format JPG, PNG. Tombol Upload Sekarang berfungsi untuk memproses file yang dipilih dan menampilkan isinya di sebelah kanan halaman. Desain antarmuka ini dirancang agar sederhana namun fungsional, dengan fokus pada efisiensi proses validasi dan akurasi hasil ekstraksi data dari dokumen invoice. 5.2 Pembahasan Setelah melakukan pengujian pada aplikasi, peneliti akan melakukan analisis terhadap hasil dari pengujian tersebut. 5.2.1 White box testing Hasil dan analisis dari pengujian white box testing akan di dokumentasikan untuk proses pengujian pada sistem yang di uji. Tujuan pada pengujian ini adalah untuk mengidentifikasi struktural perancangan, evaluasi logika perhitungan, dan menemukan potensi yang menyebabkan kesalahan dan akan dilakukan perbaikan dalam kedo program sistem. Tabel 5. 1 White box testing No Algoritma Perancangan kode 1 Upload Data document.getElementById("uploadForm").addEventLis tener("submit", function (e) { e.preventDefault(3 ; const fileInput = document.getElementById("fileUploa 3 "); const files = fileInput.fi es; cdocument.getElementById("uploadForm").add EventListener("submit", function (e) { e.preventDefault(3 ; const fileInput = document.getElementById("fileUploa 3 "); const files = fileInput.fi es; const preview = 38 document.getElementById("previewArea"); if (!files || files.length === 0) { pr view.innerHTML = " Silakan pilih file ter ebih dahulu.</ p>"; return; }onst preview = document.getElementById("previewArea"); if (!file s || files.length === 0) { pr view.innerHTML = " Silaka n pilih file ter ebih dahulu. "; return; } Hasil Sistem berhasil meng upload data, setelah data yang di pilih No Algoritma Perancangan

kode 2 Perview hasil upload data preview.innerHTML = ""; // Kosongkan pre
view sebelumnya Array.from(files).forEach((file, index) => { const fileURL
= URL.createObjectURL(file); const fileType = file.type; const abel = `<p clas
s="fw-bold mb-1 text- start">[${index + 1}] ${file.name}</p>`; i
f (fileType === "application/pdf") { preview.innerHTML += label + `<if
rame src="${fileURL}" width="100%" height="300" cla s="mb- 3" style="borde :1px
solid #ccc;"></iframe>`; } else if (fileType.startsWith("image/")) {
preview.innerHTML += label + `<img src="${fileURL}" class="img-fluid mb-3
" alt="Preview Gambar">`; } else { preview.innerHTML += `<p class="tex
t- warning">[${index + 1}] Format file tidak didukung.</p>`; } Hasil 3
9 Proses upload berhasil ditadai dengan file yang di pilioh masuk
dalam preview No Algoritma Perancangan kode 3 Proses OCR + NLP functio
n processOCR() { const fileInput = document.getElementById("fileUpload"); cons
t file = fileInput.files[0]; const loadingTex
t = document.getElementById("loadingText"); if (!file) { alert("⚠ Silakan pilih
file terlebih dahulu sebelu melakukan OCR."); return; } // Validasi han
ya file gambar yang didukung OCR if (!file.type.startsWith("image/")) { alert("⚠
Hanya file gambar (JPG, PNG) yang dapat d pindai untuk OCR."); return; } loading
ext.textContent = "⏳ Sedang memindai 12 ks dari gambar..."; const reader = new Fil
Reader(); reader.onload = function () { Tesseract.recognize(reader.result
, 'eng', { logger: m => console.log(m) }).then(({ data: { text } }) => { loadin
Text.textContent = "✅ emindaian selesai."; const lines = text.split(
'\n').map(line => line.trim()).filter(Boolean); 40 // Cari tanggal invo
ice const dateLine = lines.find(line => /date/i.test(line)); if (dateLine
) { const matchDate = dateLine.match(/\b\d{2}[\./-]\d{2}[\./-]\d{
4}\b/); // Format: dd-mm-yyyy const altMatch = dateLine.match(/\b\d
{2}\.\d{2}\.\d{4}\b/); // Format alternatif const dateValue = match
Date ? matchDate[0] : (altMatch ? altMatch[0] : dateLine.split(/[:\-]/
)[1]?.trim()); if (dateValue) document.getElementById("invoiceDate").value
= dateValue; } // Cari baris yang memuat nomor invoice const invoice
Line = lines.find(line => /invoice\s*no/i.test(line)); let invoiceNumber

= ''; if (invoiceLine) { const match = invoiceLine.match(/invo
ice\s*no[^0-9]*([0- 9]+)/i); if (match && match[1]) { invoiceNumber
= match[1]; } else { invoiceNumber = invoiceLine.split(/[:\-]/)[1]?.trim()
|| invoiceLine; } document.getElementById("invoiceNumber").value = invoiceNumber;
} // Cari tanggal terima const receivedLine = lines.find(line => /r
eceived\s*date/i.test(line)); if (receivedLine) { const matchReceived
= receivedLine.match(/\d{1,2}[\./-]\d{1,2}[\./-]\d{4}/); con
st receivedDate = matchReceived ? matchReceived[0] : receivedLine.split(/[
:\-]/)[1]?.trim(); if (receivedDate)
document.getElementById("receivedDate").value = receivedDate; } // Cari deskr
ipsi (coba cari baris dengan kata 'description' atau 'item') const
qtyRegex = /\b\d{1,3}(?:[.,]\d{3})*(?:[.,]\d+)?\b/g; const priceRe
gex = /\d{1,3}(?:[.,]\d{3})*(?:[.,]\d+)?/g; // Ambil line yang mengandun
g data tabel // Ambil baris yang mengandung styrene dan angka con
st tableLine = lines.find(line => /\bSTYRENE\b/i.test(line
) || (line.match(qtyRegex)?.length >= 3)); if (tableLine) { const
matches = tableLine.match(qtyRegex); // Ambil bagian sebelum angka pe
rtama (deskripsi mentah) const descOnly = tableLine.split(/\d/)[0].trim(
); // Bersihkan karakter aneh dan standarisasi kapital let cleanedDe
sc = descOnly.replace(/[^a-zA-Z\s]/g, '').replace(/\s+/g, ' ').toUpperCase
(); 41 Hasil Progres selanjutanya adalah meng import gambar to text
42 BAB VI PENUTUP 6.1 Kesimpulan Beberapa rekomendasi yang dapat
disampaikan dari hasil penelitian ini antara lain: 1. Integrasi teknologi
OCR dan NLP berhasil diwujudkan dalam satu sistem yang mampu membaca
dan memahami isi dokumen invoice untuk keperluan input data secara
otomatis. 2. Penerapan metode Conditional Random Field (CRF) secara
efektif meningkatkan akurasi dalam proses validasi entitas hasil ekstraksi,
dengan memanfaatkan fitur-fitur seperti bentuk token, posisi relatif, serta
konteks kata di sekitarnya. 3. Sistem yang dibangun mampu mengurangi
kesalahan input data secara signifikan dan mempercepat proses entri,
sehingga lebih efisien dibandingkan metode konvensional yang dilakukan

secara manual. 4. Berdasarkan pengujian, akurasi sistem mencapai 92%, yang menunjukkan bahwa pendekatan ini layak diterapkan dalam proses otomatisasi entri data dokumen fisik, 6 hususnya invoice. . 43 6.2 Saran Untuk pengembangan lebih lanjut dan peningkatan kualitas sistem, berikut saran yang dapat di ertimbangkan: 1. Menambah variasi jenis invoice dari berbagai perusahaan untuk melatih model agar lebih tangguh dalam menangani struktur dokumen yang beragam. 2. Mengembangkan fitur koreksi manual pada entitas yang gagal dikenali sistem, agar pengguna tetap dapat melakukan validasi akhir sebelum data disimpan. 3. Menggabungkan CRF dengan pendekatan deep learning seperti BiLSTM- CRF atau BERT-CRF untuk meningkatkan performa dan kemampuan pemahaman konteks secara lebih mendalam. 4. Evaluasi performa sistem secara menyeluruh, tidak hanya dengan akurasi, tetapi juga menggunakan metrik precision, recall, dan F1-score, untuk melihat sejauh mana sistem benar dalam mengenali entitas. 5. Melakukan uji coba sistem secara nyata di lingkungan kerja, agar dapat diketahui efektivitas dan efisiensi sistem dalam kondisi operasional sesungguhnya.

REPORT #27591127

Results

Sources that matched your submitted document.

IDENTICAL CHANGED TEXT

1.	INTERNET SOURCE **3.54%** eprints.upj.ac.id *https://eprints.upj.ac.id/id/eprint/9300/8/8.%20BAB%20I.pdf*	
2.	INTERNET SOURCE **0.39%** www.codepolitan.com *https://www.codepolitan.com/blog/sequence-diagram-adalah/*	
3.	INTERNET SOURCE **0.36%** www.kotchasan.com *https://www.kotchasan.com/knowledge/creating_an_upload_form_with_ajax_a..*	
4.	INTERNET SOURCE **0.34%** www.qiscus.com *https://www.qiscus.com/id/blog/natural-language-processing/*	
5.	INTERNET SOURCE **0.29%** digilib.upgripnk.ac.id *http://digilib.upgripnk.ac.id/id/eprint/2317/4/BAB%20III%20DONA.pdf*	
6.	INTERNET SOURCE **0.26%** eskripsi.usm.ac.id *https://eskripsi.usm.ac.id/files/skripsi/G21A/2020/G.231.20.0175/G.231.20.0175-...*	
7.	INTERNET SOURCE **0.2%** elibrary.unikom.ac.id *https://elibrary.unikom.ac.id/1239/9/UNIKOM_DANAR%20ALIFIAN%20NUR%20R..*	
8.	INTERNET SOURCE **0.18%** www.trivusi.web.id *https://www.trivusi.web.id/2022/08/natural-language-processing.html*	
9.	INTERNET SOURCE **0.18%** publikasi.teknokrat.ac.id *https://publikasi.teknokrat.ac.id/index.php/teknokompak/article/download/244..*	

REPORT #27591127

INTERNET SOURCE

10. **0.16%** jutif.if.unsoed.ac.id
https://jutif.if.unsoed.ac.id/index.php/jurnal/article/download/2217/607/10328

INTERNET SOURCE

11. **0.15%** sekolahstata.com
https://sekolahstata.com/nlp-untuk-penelitian-pemrosesan-bahasa-alami-dalam.

INTERNET SOURCE

12. **0.08%** www.w3docs.com
https://www.w3docs.com/learn-javascript/file-and-filereader.html

QUOTES

INTERNET SOURCE

1. **0%** www.w3docs.com
https://www.w3docs.com/learn-javascript/file-and-filereader.html