EKSTRAKSI ENTITAS DOKUMEN MULTI-FORMAT MENGGUNAKAN *OPTICAL CHARACTER RECOGNITION* DAN *NAMED ENTITY RECOGNITION*

SEMINAR BIDANG KAJIAN

ALI ISRA
99223101

PROGRAM DOKTOR TEKNOLOGI INFORMASI
UNIVERSITAS GUNADARMA
JUNI 2024

Daftar Isi

Bab 1

Pendahuluan

1.1 Latar Belakang

Penerapan teknologi informasi menjadi hal yang krusial dalam mengoptimalkan kinerja organisasi dalam berbagai bidang, baik di dunia industri yang berorientasi pada bisnis maupun di institusi pemerintah. Transformasi digital menjadi prioritas dalam upaya peningkatan percepatan pencapaian tujuan dari setiap organisasi tersebut. Dalam bidang pemerintahan, Pemerintah Repuplik Indonesia menjadikan transformasi digital sebagai salah satu kebijakan prioritas dalam upaya peningkatan penyelenggaraan pelayanan publik. Tujuan dari kebijakan ini adalah mempermudah akses, peningkatan transparansi dan percepatan proses pelayanan publik. Kebijakan ini telah ditetapkan dalam Undang-undang Nomor 25 Tahun 2009 tentang Pelayanan Publik yang diturunkan dalam Peraturan Presiden Republik Indonesia Nomor 95 Tahun 2018 tentang Sistem Pemerintahan Berbasis Elektronik. Hal ini juga sejalan dengan Instruksi Presiden Republik Indonesia Nomor 3 Tahun 2003 tentang Kebijakan dan Strategi Nasional Pengembangan E-Goverment. Kebijakan-kebijakan ini memiliki tujuan utama menciptakan tata kelola pemerintahan yang baik (good governance) yang efektif, efisien serta transparan.

Penggunaan teknologi informasi sudah merambah kepada penerapan kecerdasan buatan (artificial intelligence). Kecerdasan buatan memiliki peran penting dalam melakukan otomatisaisi proses pelayanan publik. Salah satu proses yang masih menjadi kendala dalam penyelenggaraan pelayanan publik di berbagai institusi pemerintah adalah proses verifikasi dan validasi yang masih manual terhadap dokumen persyaratan pengajuan layanan publik. Penerapan kecerdasan buatan sangat dibutuhkan untuk mengenali entitas data yang terkandung dalam dokumen dalam bentuk teks gambar, sehingga proses verifikasi secara otomatis dapat dilakukan dengan membandingkan data tersebut ke

database tertentu sebagai rujukan validasi. Bahkan entitas ini dapat digunakan dalam otomatisasi pengisian form (autofill) pengajuan layanan. Optical Character Recognition (OCR) adalah algoritma yang dapat melakukan konversi teks gambar ke dalam bentuk teks digital, sedangkan Large Language Model (LLM) memiliki kemampuan analisis semantik untuk mengidentifikasi entitas data yang diperlukan. Kombinasi kedua metode ini memiliki peranan yang penting dalam mengotomatisasi proses verifikasi dokumen maupun penginputan otomatis dari hasil ekstraksi entitas suatu dokumen. Hal ini juga dapat mempercepat proses yang dilakukan dan meningkatkan keakuratan dalam penyelenggaraan pelayanan publik utamanya penanganan pada volume dokumen yang besar dari jumlah ajuan layanan yang besar. Sebagai contoh, berdasarkan data SIPINTER dan SIJAFUNG Lembaga Layanan Pendidikan Tinggi (LLDIKTI) Wilayah IX, pada tahun 2022 hingga 2023 terdapat 30.322 dokumen dengan rata-rata 12.716 dokumen pertahun yang harus diverifikasi untuk 56 pelayanan publik yang diselenggarakan oleh LLDIKTI Wilayah IX. Dokumen yang diverifikasi berupa ijazah, KTP, surat keputusan maupun dokumen-dokumen terkait persyaratan layanan yang memuat teks berupa entitas seperti nama, tanggal lahir, program studi, angka kredit yang dapat dikenali melalui penerapan NER menggunakan LLM. Pendekatan ini diharapkan mampu meningkatkan transparasi dan efisiensi waktu proses verifikasi serta mampu mengurangi kesalahan manusia dalam proses verifikasi. Penerapan teknologi diharapkan dapat ikut mendukung kebijakan peningkatan pelayanan publik yang handal sehingga dapat memperkuat integritas data serta mampu meningkatkan kepercayaan publik terhadap institusi pemerintah.

Dalam proses ekstraksi dokumen teks gambar hasil scan kedalam bentuk teks digital, berbagai pendekatan menggunakan algoritma OCR telah dilakukan melalui penelitian untuk meningkatkan akurasi, mengurangi waktu pemrosesan, dan mengatasi kompleksitas dokumen. [Firhan Maulana Rusli et al., 2020] mengimplementasikan OCR dengan post-processing menggunakan NLP untuk ekstraksi data dari kartu identitas Indonesia, mencapai F-score 0.78 dengan waktu pemrosesan 4510 milidetik per kartu. [Rifiana Arief et al., 2018] menggunakan Google Vision OCR dalam lingkungan Apache Hadoop untuk ekstraksi teks dari dokumen skala besar, mencapai akurasi 100% dengan waktu ekstraksi dua kali lebih cepat dibandingkan ekstraksi manual. [Vedant Kumar et al., 2020] mengatasi masalah watermark dan bayangan pada gambar tagihan menggunakan OpenCV sebelum ekstraksi teks dengan Tesseract OCR, menunjukkan efektivitas dalam mengelola gambar yang kurang ideal. [A. Ceniza et al., 2018] mengembangkan aplikasi mobile untuk mengenali teks dalam gambar dokumen yang terdegradasi menggunakan Tesseract dengan Binarisasi Gambar Dokumen Adaptif, mencapai akurasi ka-

rakter rata-rata 93.17% dan akurasi kata 85.82%. Metode ini menunjukkan bagaimana teknologi OCR dapat disesuaikan untuk berbagai kondisi dokumen dan kebutuhan pengolahan, meskipun tantangan seperti biaya komputasi tinggi dan kebutuhan untuk pelatihan data yang cukup tetap menjadi pertimbangan penting dalam pengembangan dan penerapan sistem ekstraksi teks.

Di sisi lain, Berbagai pendekatan menggunakan LLM dalam ekstraksi entitas dokumen juga terus dikembangkan melalui penelitian-penelitian yang menggunakan metode dan algoritma yang berbeda dengan kelebihan dan keterbatasan masing-masing dalam hal akurasi, kompleksitas, waktu pemrosesan, serta biaya. [Perot et al., 2023] memperkenalkan metodologi LMDX yang menyesuaikan LLM untuk ekstraksi informasi dokumen dengan menggabungkan pengkodean tata letak dan mekanisme grounding. [Wu et al., 2024] membahas Ekstraksi Entitas Terstruktur (SEE) menggunakan LLM dan memperkenalkan metrik AESOP untuk evaluasi kinerja, menunjukkan peningkatan dalam efisiensi dan efektivitas. [Huang et al., 2021] mengusulkan kerangka kerja E2GRE, yang menggunakan mekanisme perhatian dan input yang dipandu entitas untuk meningkatkan ekstraksi hubungan dan prediksi bukti tingkat dokumen, mencapai kinerja terdepan di dataset DocRED tetapi dengan peningkatan kompleksitas karena kerangka pelatihan bersama.

Penelitian tentang Optical Character Recognition (OCR) dan Large Language Models (LLMs) telah menunjukkan kemajuan yang signifikan, namun masih terdapat celah yang dapat ditangani melalui investigasi lebih lanjut. Dalam ranah OCR, meskipun telah ada kemajuan dalam ekstraksi teks dari gambar, tantangan masih ada dalam menangani dokumen multi-format yang bervariasi dalam tata letak dan struktur, seperti elemen teks dan non-teks yang bercampur, atau format dokumen yang berbeda [Firhan Maulana Rusli et al., 2020, Kreshnik Vukatana, 2022]. Hal ini menunjukkan kebutuhan akan sistem OCR yang lebih canggih yang mampu menangani berbagai format dokumen dengan akurasi tinggi. Di sisi lain, meskipun LLM seperti BERT telah berhasil diterapkan untuk Named Entity Recognition (NER) dalam berbagai bahasa, penerapannya dalam konteks dokumen multi-format, terutama yang melibatkan bahasa dengan sumber daya terbatas, masih kurang dieksplorasi [Kryeziu and Shehu, 2023]. Ini menunjukkan peluang untuk mengintegrasikan teknologi OCR dengan kemampuan LLM yang maju untuk meningkatkan ekstraksi entitas dari dokumen multi-format. Oleh karena itu, topik yang diusulkan dalam penelitian ini adalah "Ekstraksi Entitas Dokumen Multi-format Menggunakan Optical Character Recognition (OCR) dan Named Entity Recognition (NER)". Tujuannya adalah untuk mengembangkan sistem yang memanfaatkan kekuatan OCR untuk ekstraksi teks yang akurat dari berbagai

format dokumen dan LLM untuk pengenalan entitas yang canggih. Integrasi ini dapat secara signifikan mengurangi kesalahan entri data manual, meningkatkan kecepatan pemrosesan, dan memperluas aplikabilitas sistem pemrosesan dokumen di berbagai bahasa dan format, memberikan solusi komprehensif yang meningkatkan aksesibilitas dan kegunaan data dalam sistem otomatis terutama pada Otomatisasi verifikasi dokumen persyaratan pada proses pengajuan pelayanan publik.

1.2 Batasan dan Tujuan

Adapun batasan masalah dari penelitian ini adalah sebagai berikut:

- Data berupa dokumen yang diperoleh dari file dokumen persyaratan pengajuan layanan publik yang diajukan oleh pengguna layanan pada aplikasi SIPINTER LLDIKTI Wilayah IX
- Entitas dokumen yang akan diekstraksi berupa informasi yang termuat dalam dokumen persyaratan pengajuan layanan publik meliputi ijazah, KTP, SK Jabatan Fungsional Dosen, Surat Pernyataan dan sejenisnya.
- Tujuan umum dari penelitian ini adalah menghasilkan sistem yang mampu melakukan ekstraksi entitas data pada dokumen dengan berbagai format sehingga data tersebut dapat dimanfaatkan dalam proses otomatisasi penginputan maupun verifikasi dokumen. Sedangkan tujuan khususnya adalah menghasilkan model yang mampu melakukan ekstraksi entitas dokumen multi-format dengan tingkat akurasi yang tinggi dan waktu proses yang lebih cepat.

1.3 Kontribusi

Adapun kontribusi sistem dari sisi keilmuan dari penelitian ini adalah

- menghasilkan model ekstraksi entitas dokumen multi-format dengan nilai akurasi yang tinggi dengan waktu proses yang lebih cepat.

Sedangkan kontribusi sistem terhadap bidang penerapan di masyarakat adalah:

- Mempermudah dan mempercepat proses penginputan dan verifikasi dokumen pada penyelenggaraan pelayanan publik
- Meningkatkan tingkat kepercayaan masyarakat terhadap penyelenggaraan pelayanan publik

Bab 2

Tinjauan Pustaka

2.1 Tinjauan 1: Metode Ekstraksi Teks Gambar menjadi Teks Digital

Serangkaian studi pada penelitian-penelitian yang membahas terkait metode untuk melakukan ekstraksi teks gambar menjadi teks digital dari berbagai tipe dokumen, dilakukan dengan menggunakan OCR (Optical Character Recognition). Terdapat berbagai penelitian yang membahas tentang penanganan ekstraksi teks gambar menjadi teks digital dengan berbagai pendekatan, berikut adalah penelitian yang telah dilakukan:

Tabel 2.1: Tabel Penelitian tentang Ekstraksi Teks

Penulis/Judul	**Tujuan Penelitian**	**Metode**	**Hasil/Temuan**
[Firhan Maulana Rusli et al., 2020] **Indonesian ID Card Extractor Using Optical Character Recognition and Natural Language Post-Processing**	mengembangkan sistem ekstraksi kartu identitas Indonesia yang efektif menggunakan OCR dan NLP	OCR (PyTesseract); NLP (RegGex,punctuation removal,word-to-number conversion,special handling utk field)	Hasil tertinggi F-score = 0.84 dengan waktu proses 4510 milliseconds per KTP.
[Hoan Tran Viet et al., 2019] **A Robust End-To-End Information Extraction System for Vietnamese Identity Cards**	Mengekstrak informasi dari kartu identitas Vietnam yang diambil oleh kamera smartphone, meskipun tantangan latar belakang yang bervariasi, perspektif, kabur, dan font yang beragam	Corner detection, national emblem detection, deep CNN; OCR architecture for text recognition	akurasi rata-rata lebih tinggi dari 91% dan mengurangi waktu pemrosesan dibandingkan dengan state of the art.
[S. Surana et al., 2022] **Text recognition for Vietnamese identity card based on deep features network**	Mengembangkan metode pengenalan karakter optik (OCR) untuk kartu identitas Vietnam untuk mendukung verifikasi identitas dalam proses Know Your Customer (KYC)	OCR; deep features network	Akurasi lebih dari 96.7% untuk karakter dan 89.7% untuk kata sesuai entitas pada ID Card.

[Kreshnik Vukatana, 2022] **OCR and Levenshtein distance as a measure of image quality accuracy for identification documents**	mengembangkan metode untuk menilai kualitas dan keterbacaan gambar, khususnya dokumen identifikasi, agar dapat meningkatkan pra-pemrosesan dataset untuk klasifikasi dokumen menggunakan AI	OCR and Levenshtein distance	memilah input data berdasarkan kualitas gambar
[Jeklin Harefa et al., 2022] **ID Card Storage System using Optical Character Recognition (OCR) on Android-based Smartphone**	mengembangkan sistem untuk mengekstrak data pribadi dari gambar KTP menggunakan OCR di platform mobile	OCR; image preprocessing, character recognition, and character extraction	Akurasi 87,57%; Rata-rata 43 kata (2.137 karakter) dikenali dari 1 ID Card
[Anurag Tiwari, 2021] **Data Extraction from Images through OCR**	meningkatkan efisiensi dan aksesibilitas manajemen dokumen melalui penggunaan teknologi OCR	Tesseract OCR Engine, TesseractJS, other JavaScript frameworks	Akurasi di atas 90% dengan kelemahan waktu proses cukup lama
[Akpinar et al., 2018] **Extracting table data from images using optical character recognition text**	mengekstrak data tabel dari dokumen berbasis gambar menggunakan OCR	Specific algorithms to extracting tabular data from OCR text	Berhasil mendeteksi rows dan columns

[Chandni Kaundilya et al., 2019] **Automated Text Extraction from Images using OCR System**	mengekstrak teks dari gambar secara otomatis	Tesseract OCR Engine: text detection, text localization, text segmentation, and binarization.	
[Rao et al., 2019] **Optical Character Recognition from Printed Text Images**	mengembangkan metode yang sederhana dan efektif untuk mengekstrak teks dari gambar dokumen, bahkan dalam keadaan adanya noise dan kabur	FAST algorithm	Akurasi $> 90\%$
[Siddharth Salar Et.al, 2021] **Automate Identification and Recognition of Handwritten Text from an Image**	mengembangkan algoritma AI untuk mengekstrak dan mengidentifikasi teks tulisan tangan dari gambar	Pytesseract OCR engine, convolutional neural networks, rectified linear units (ReLUs), and pooling layers	Akurasi $> 95\%$
[Karanrat Thammarak et al., 2022] **Comparative analysis of Tesseract and Google Cloud Vision for Thai vehicle registration certificate**	Mengekstrak informasi dari sertifikat registrasi kendaraan Thailand	Convolutional neural network (CNN), recurrent neural networks (RNNs), and long short-term memory (LSTM)	Akurasi 84,43%

[Neha Agrawal and Arashdeen Kaur, 2018] **An Algorithmic Approach for Text Recognition from Printed/Typed Text Images**	mengembangkan algoritma untuk mengekstrak teks secara akurat dari gambar dokumen yang telah discan	Otsu's algorithm for segmentation and Hough transform method for skew detection	Akurasi 93%
[Bektemyssova Gulnara and Akhmer Yerassyl, 2022] **Using Image Processing and Optical Character Recognition to Recognise ID cards in the Online Process of Onboarding**	mengembangkan metode yang efektif untuk mengenali dan memverifikasi kartu ID elektronik menggunakan pemrosesan gambar dan OCR	image Detection; OCR	Hasil Deteksi Electronic ID Card 99%
[Sujata Desai et al., 2020] **An approach for Text Recognition from Document Images**	Cara efektif mengekstrak teks dari dokumen yang telah discan dan gambar	Otsu segmentation algorithm dan Hough transforming method ; OCR untuk penegnalan karakter	Akurasi 93%

[Rifiana Arief et al., 2018] **Automated Extraction of Large Scale Scanned Document Images using Google Vision OCR in Apache Hadoop Environment**	menemukan cara efisien mengekstrak teks dari gambar dokumen yang telah discan dalam skala besar	general text extraction pipeline of preprocessing, detection, localization, extraction, enrichment, and OCR, as well as the use of the Google Vision OCR algorithm	Akurasi 100%

Studi yang dilakukan oleh [Firhan Maulana Rusli et al., 2020] yaitu menggabungkan OCR dengan NLP (Natural Language Processing) untuk meningkatkan akurasi ekstraksi teks pada KTP Indonesia. Dari 50 gambar KTP Indonesia yang terdiri dari 25 gambar diambil langsung menggunaka kamera dan 25 gambar lainnya adalah hasil pindai menggunakan scanner. Melalui proses preprosesing yaitu dengan mengubah gambar menjadi grayscale sebelum menjadi format binner, Ekstraksi teks ke dalam bahasa Indonesia menggunakan Library Python Tesseract menggunakan parameter bahasa "ind". Proses selanjutnya adalah melakukan perbaikan terhadap ketidaksempurnaan OCR dalam menghasilkan karakter seperti angka '0' yang ditulis sebagai huruf 'O', ataupun angka '1' ditulis sebagai huruf 'l'. Setelah itu dilanjutkan dengan mengekstraksi konten dari setiap entitas dengan memisahkan kalimat berdasarkan tanda titik dua. Proses ini dilakukan dengan menggunakan teknik NLP yaitu Reguler Expression (RegEx). Hasil dari peneltian ini adalah F-score secara keseluruhan sebesar 0.78 dengan kecepatan proses rata-rata per KTP mencapai 4510 milidetik. Setiap entitas dilengkapi dengan Skor Confidence untuk tingkat kepercayaan terhadap tiap entitas.

Studi yang dilakukan oleh [Duc Phan Van Hoai et al., 2021] yaitu melakukan pengenalan teks menggunakan OCR pada kartu identitas Vietnam. Dataset yang digunakan adalah 2500 gambar kartu identitas Vietnam. Tahap preprosesing yang dilakukan sama dengan penelitian sebelumnya, yaitu dengan melakukan perubahan gambar ke mode grayscale yang dilanjutkan dengan tahap deteksi bounding box dari teks yang ditemukan menggunakan CTPN (Connectionist Text Proposal Network). Selanjutnya, untuk mendeteksi teks dalam area bounding box, model menggunakan CRNN (Convolutional Recurrent Neural Network) dengan GRU (Gated Reccurent Units). Hasil dari penelitian ini adalah akurasi dalam pengenalan mencapai 96.7% dengan CER (Character Error Rate) sebesar 8% dan WER (Word Error Rate) sebesar 9%.

2.2 Tinjauan 2: Metode Ekstraksi Entitas dari Dokumen Teks

Penelitian terkait dengan ekstraksi entitas suatu dokumen pada umumnya memanfaatkan kapasitas Large Language Model (LLM) yang memiliki kemampuan yang luas dalam pemahaman bahasa. Berikut adalah penelitian-penelitian yang membahas metode pengenalan entitas dokumen menggunakan LLM:

Tabel 2.2: Tabel Penelitian tentang Ekstraksi Teks

Penulis/Judul	**Tujuan Penelitian**	**Metode**	**Hasil/Temuan**
[Shi et al., 2019] **Entity Relationship Extraction Based on BLSTM Model**	mengembangkan model yg efektif dapat mengekstrak hubungan antar entitas dari teks bahasa alami menggunakan Deep Learning	BLSTM (Bidirectional LSTM) model and the Bidirectional LSTM model with Attention	Meningkatkan jumlah diemensi kata dan keterhubungan ekstraksi
[Monal Deshmukh and S. Maheshwari, 2019] **Free Form Document Based Extraction Using ML**	mengembangkan metode untuk mengekstrak informasi secara otomatis dan akurat dari dokumen teks bebas, khususnya dokumen yang telah discan dan gambar	Supervised learning algorithms, tokenization, POS tagging, entity detection, and dependency parsing	Hasil ekstrak digunakan untuk otomasi pengisian form
[Huang et al., 2021] **Entity and Evidence Guided Document-Level Relation Extraction**	mengembangkan kerangka kerja yang dapat secara bersamaan mengekstrak hubungan dan kalimat bukti yang mendukung hubungan tersebut pada level dokumen	E2GRE	Top 1 untuk evidence prediction berdasarkan SOTA
[Perot et al., 2023] **LMDX: Language Model-based Document Information Extraction and Localization**	Mengembangkan LLM adaptif untuk ekstraksi informasi dokumen semiterstruktur berkualitas tinggi, dan mengatasi tantangan pengkodean tata letak dan membumikan informasi yang diekstrak	LMDX, a methodology for adapting large language models to perform document information extraction	enables the creation of high-quality, data-efficient parsers

[Wu et al., 2024] **Structured Entity Extraction Using Large Language Models**	mengembangkan metode yang lebih efektif dan efisien untuk mengekstrak informasi terstruktur dari teks yang tidak terstruktur menggunakan LLM	Structured Entity Extraction (SEE), Approximate Entity Set OverlaP (AESOP) metric, decomposes the extraction task into multiple stages using LLM	menemukan decomposes the extraction task into multiple stages using LLM
[Devlin et al., 2019] **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**	mengembangkan model representasi bahasa yang dapat secara efektif menangkap konteks dua arah	masked language model (MLM) pre-training task, where 15% of input tokens are randomly masked and the model is trained to predict those masked tokens	improvement over the ESIM+ELMo baseline and 8.3% over OpenAI GPT on the SWAG task - BERT LARGE obtains a GLUE score of 80.5, compared to 72.8 for OpenAI GPT
[Wang et al., 2017] **Named Entity Recognition with Gated Convolutional Neural Networks**	mengembangkan model neural network yang lebih baik untuk NER yang dapat mengatasi tantangan data pelatihan yang terbatas dan distribusi entitas bernama yang jarang	Gated convolutional neural networks (GCNN) for feature extraction, Conditional random field (CRF) for sequence prediction	

[Zhang and Yang, 2018] **Chinese NER Using Lattice LSTM**	meningkatkan kinerja NER dalam bahasa Mandarin dengan lebih baik memanfaatkan informasi kata, tanpa bergantung pada segmentasi kata yang rentan terhadap kesalahan	Lattice LSTM	F1-score 71.62%
[Kryeziu and Shehu, 2023] **Bert Based Named Entity Recognition for the Albanian Language**	mengembangkan sistem NER yang lebih baik untuk bahasa Albania	BERT (Bidirectional Encoder Representations from Transformers) - mbert-base-albanian-cased-ner (a BERT model fine-tuned on the alb-dataset for NER in Albanian) - Other BERT-based models like BERT-base, BERT-large, multilingual BERT, and AraBERT	Meningkatkan akurasi ekstraksi entitas person, organisasi, lokasi dari teks Albania

2.3 Perbandingan Tinjauan

Pada tinjauan penelitian yang membahas tentang metode ektraksi teks dari dokumen gambar, beberapa metode yang dapat diadopsi dalam penanganan ekstraksi teks dari dokumen gambar multi-format antara lain:

1. Penggunaan Pytesseract
   Pytesseract, sebagai implementasi Python dari Tesseract OCR, telah menunjukkan performa yang baik dalam penelitian oleh [Firhan Maulana Rusli et al., 2020]. Pytesseract menawarkan fleksibilitas dalam pengolahan bahasa dan memiliki kemampuan untuk menangani berbagai format dokumen dengan akurasi yang relatif tinggi. Ini membuatnya cocok untuk aplikasi yang memerlukan ekstraksi teks dari dokumen multi-format.

2. Deep Learning dan Neural Networks
   Penerapan jaringan saraf dalam OCR, seperti yang dibahas oleh [Duc Phan Van Hoai et al., 2021] dan [Arora et al., 2020], menunjukkan peningkatan signifikan dalam akurasi pengenalan teks. Model-model ini, khususnya yang menggunakan arsitektur deep convolutional neural networks (CNN) dan recurrent neural networks (RNN), sangat efektif dalam memahami konteks dan nuansa dalam gambar dokumen yang kompleks. Penerapan teknik ini dapat meningkatkan kemampuan model OCR untuk secara akurat mengenali dan mengekstrak teks dari berbagai jenis dokumen.

3. Preprocessing Techniques Teknik pra-pemrosesan yang efektif sangat penting untuk meningkatkan kinerja OCR. Seperti yang diilustrasikan dalam penelitian oleh [Rifiana Arief et al., 2018] dan [A. Ceniza et al., 2018], teknik seperti binarisasi adaptif dan peningkatan kontras dapat mempersiapkan gambar dokumen dengan lebih baik untuk pengolahan OCR, mengurangi kesalahan pengenalan, dan meningkatkan akurasi keseluruhan.

4. Post-Processing untuk Koreksi Kesalahan Seperti yang dijelaskan oleh [Firhan Maulana Rusli et al., 2020], penggunaan alat-alat NLP untuk pasca-pemrosesan teks yang diekstraksi dapat membantu memperbaiki kesalahan yang umum terjadi dalam hasil OCR. Teknik-teknik ini dapat meliputi penghapusan tanda baca yang salah, koreksi ejaan, dan penyesuaian format teks untuk meningkatkan keakuratan dan keterbacaan output OCR.

Sedangkan pada tinjauan penelitian yang membahas metode untuk melakukan ekstraksi entitas dari kumpulan teks dari suatu dokumen yang dapat diadopsi dalam pe-

ngembangan metode untuk penanganan ekstraksi entitas pada dokumen multi-format antara lain:

1. Adaptasi Model Bahasa untuk NER
   Adaptasi model bahasa besar seperti BERT untuk tugas NER adalah tema yang berulang. [Kryeziu and Shehu, 2023] mengeksplorasi penggunaan BERT multibahasa, yang ditajamkan pada korpus bahasa Albania untuk NER, menyoroti fleksibilitas model di berbagai bahasa dan efektivitasnya dalam mengekstrak entitas bernama dari teks. Pendekatan ini menekankan potensi pembelajaran transfer dan adaptabilitas model yang telah dilatih sebelumnya untuk tugas NER spesifik, yang dapat sangat penting untuk menangani dokumen dalam berbagai bahasa atau format.

2. Ekstraksi Berbasis Entitas dan Bukti
   [Huang et al., 2021] memperkenalkan kerangka kerja baru, E2GRE, yang menggabungkan ekstraksi entitas dan bukti untuk meningkatkan ekstraksi hubungan tingkat dokumen. Metode ini menggunakan mekanisme perhatian untuk fokus pada bagian teks yang relevan, meningkatkan akurasi ekstraksi hubungan entitas. Teknik ini dapat diadaptasi untuk meningkatkan presisi ekstraksi entitas dalam dokumen multi-format dengan fokus pada segmen teks yang relevan, sehingga meningkatkan kualitas keseluruhan proses ekstraksi.

3. Ekstraksi Entitas Terstruktur Menggunakan LLM [Wu et al., 2024] membahas penggunaan Model Bahasa Besar (LLMs) untuk ekstraksi entitas terstruktur, menekankan pada dekomposisi tugas menjadi beberapa tahap untuk meningkatkan efisiensi dan efektivitas. Pendekatan bertahap ini bisa sangat bermanfaat untuk dokumen multi-format di mana bagian atau bagian yang berbeda mungkin memerlukan strategi ekstraksi yang berbeda.

4. Pemanfaatan Jaringan Saraf Konvolusional (CNN) [Wang et al., 2017] mengusulkan arsitektur CNN berbasis gerbang untuk NER, yang dicatat karena efisiensi pelatihannya dan kemampuannya untuk menangani berbagai bahasa. Penggunaan CNN bisa sangat efektif untuk dokumen multi-format karena dapat memproses fitur spasial dalam dokumen terstruktur (seperti formulir atau tabel) di mana tata letak memainkan peran penting dalam memahami konten.

5. Penggabungan Struktur Lattice untuk Skrip Kompleks [Zhang and Yang, 2018] mengeksplorasi LSTM lattice untuk NER Cina, yang secara efektif menangani kompleksitas skrip Cina dengan mengintegrasikan informasi tingkat karakter dan

kata. Metode ini dapat diadaptasi untuk dokumen yang berisi tata letak atau skrip kompleks, memastikan bahwa kekayaan semantik dari format tersebut ditangkap dengan memadai.

6. Ekstraksi dan Lokalisasi Informasi Dokumen [Perot et al., 2023] memperkenalkan LMDX, metodologi untuk menyesuaikan LLM untuk ekstraksi informasi dokumen yang mencakup lokalisasi entitas. Pendekatan ini sangat penting untuk dokumen multi-format karena memungkinkan ekstraksi dan penempatan entitas yang tepat, yang sangat penting untuk dokumen di mana format dan struktur menyampaikan makna.

Bab 3

Metodologi

Intinya memberikan ringkasan usulan metodologi / pendekatan / arsitektur yang digunakan untuk pemecahan dari perumusan dan batasan masalah.

3.1 Motivasi

Motivasi untuk mengembangkan sistem ekstraksi entitas pada dokumen multi-format menggunakan integrasi Optical Character Recognition (OCR) dan Large Language Models (LLMs) seperti Named Entity Recognition (NER) sangat didorong oleh kebutuhan untuk meningkatkan efisiensi dan akurasi dalam pengolahan informasi dari berbagai jenis dokumen. Dalam konteks ini, OCR digunakan untuk mengonversi dokumen cetak atau tulisan tangan menjadi teks digital yang dapat diproses lebih lanjut, sementara LLMs digunakan untuk mengidentifikasi dan klasifikasi entitas dalam teks yang diekstraksi [Firhan Maulana Rusli et al., 2020, Kryeziu and Shehu, 2023]. Beberapa motivasi dari sisi implementasi adalah:

1. penggunaan OCR memungkinkan otomatisasi entri data yang sebelumnya dilakukan secara manual, yang sering kali memakan waktu dan rentan terhadap kesalahan. Dengan mengotomatiskan proses ini, organisasi dapat mengurangi waktu dan biaya yang terkait dengan entri data manual serta meningkatkan kecepatan pemrosesan dokumen [Firhan Maulana Rusli et al., 2020].

2. penerapan LLMs dalam NER membantu dalam pengenalan entitas yang akurat seperti nama, lokasi, dan tanggal dari teks yang diekstraksi. Ini sangat penting dalam banyak aplikasi seperti pengelolaan dokumen hukum, medis, dan bisnis di mana pengidentifikasian informasi yang tepat sangat kritis [Kryeziu and Shehu, 2023].

3. integrasi OCR dan LLMs dalam sistem ekstraksi entitas menawarkan kemampuan untuk menangani dokumen dalam berbagai format dan bahasa, termasuk bahasa dengan sumber daya terbatas. Ini memperluas cakupan aplikasi teknologi ini ke lebih banyak konteks dan penggunaan global, terutama dalam mendukung pengolahan bahasa yang kurang terwakili [Kryeziu and Shehu, 2023].

4. peningkatan dalam teknologi OCR dan LLMs, seperti yang ditunjukkan dalam penelitian terbaru, menjanjikan peningkatan lebih lanjut dalam akurasi dan keandalan sistem ekstraksi entitas. Ini membuka peluang untuk inovasi lebih lanjut dan aplikasi praktis dari teknologi ini dalam skenario dunia nyata
[Firhan Maulana Rusli et al., 2020, Kryeziu and Shehu, 2023].

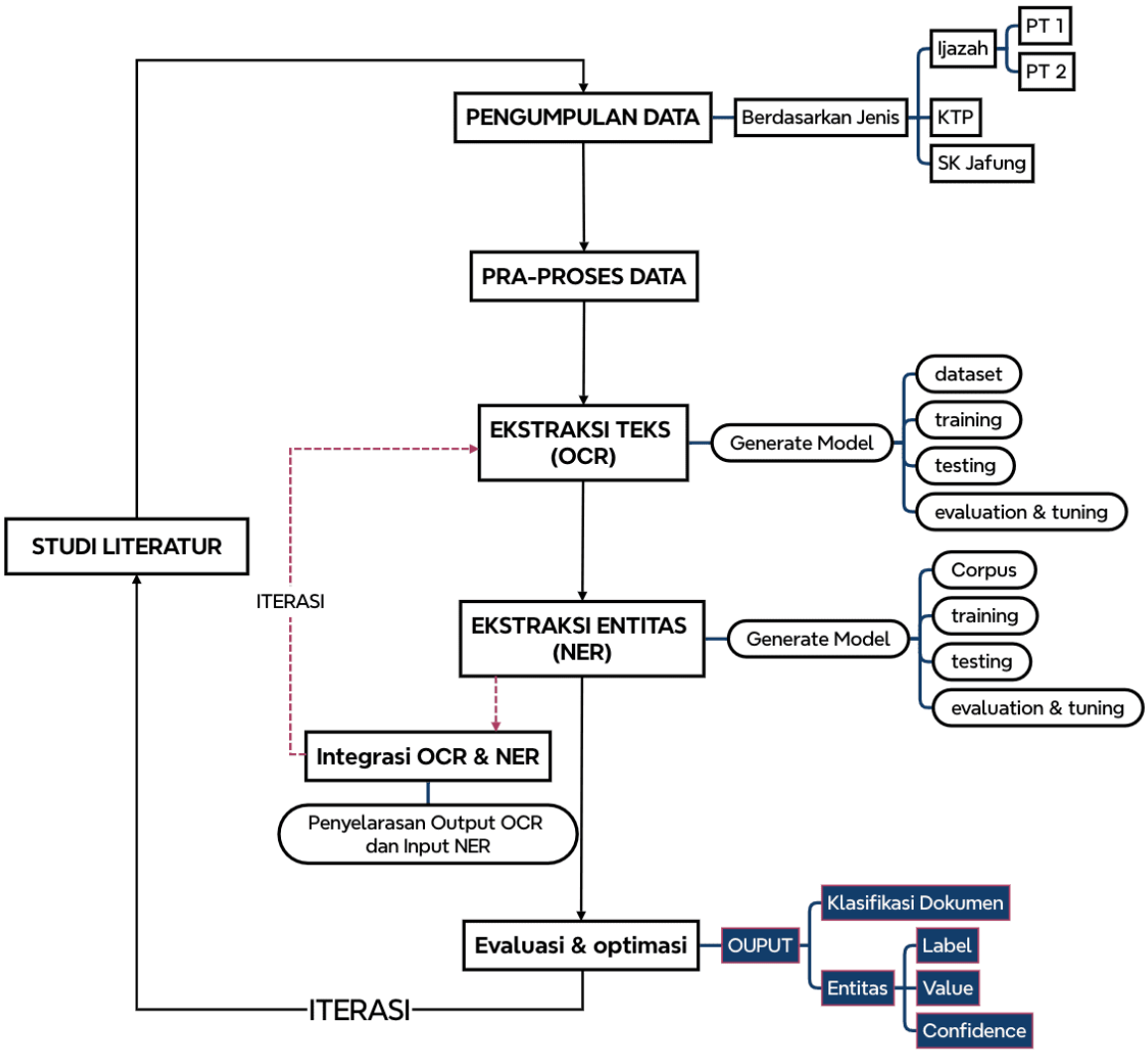
Oleh karena itu, motivasi utama di balik penelitian ini adalah untuk mengembangkan solusi yang lebih efisien dan akurat untuk ekstraksi entitas otomatis dari dokumen multi-format, yang dapat secara signifikan meningkatkan pengolahan informasi dan manajemen pengetahuan dalam berbagai sektor industri.

3.2 Framework Riset

Framework riset untuk mengembangkan sistem ekstraksi entitas pada dokumen multi-format menggunakan integrasi Optical Character Recognition (OCR) dan Large Language Models (LLMs) seperti Named Entity Recognition (NER) dapat dirancang dengan beberapa komponen utama berikut:

1. Studi Literatur
   Tahap pertama yang dilakukan adalah melakukan studi literatur terhadap berbagai penelitian untuk menelaah berbagai pendekatan dan perkembangan teknologi OCR dalam menangani pemrosesan teks gambar menjadi teks digital dan NER dalam mengklasifikasi entitas dalam teks. Selain kedua metode ini, dalam tahap ini juga . Demikian pula terkait persiapan dan pemrosesan data serta metode evaluasi yang akan digunakan.

2. Pengumpulan dan Persiapan Data
   Data berupa dokumen multi-format pada penelitian ini menggunakan dokumen yang dikumpulkan dari laman SIPINTER sebagai super-apps pelayanan publik LLDIKTI Wilayah IX. Dokumen-dokumen akan diklasifikasikan berdasarkan jenisnya dan akan dilakukan pelabelan sesuai dengan kebutuhan pada tahapan selanjutnya.



Gambar 3.1: Framework Penelitian

3. Pra-pemrosesan Dokumen
   Setelah data terkumpul, dilakukan pra-pemrosesan dokumen yang mencakup pembersihan (cleansing) dari noise, pengubahan skala dan orientasi, segmentasi, koneversi ke grayscakle atau binner dan berbagai proses pengolahan gambar untuk memudahkan proses pada tahap selanjutnya.

4. Ekstraksi Teks menggunakan OCR
   Tahap berikutnya adalah ekstraksi teks menggunakan OCR, di mana dokumen yang telah diproses di-scan dan karakternya dikenali menggunakan library OCR seperti Tesseract, yang mengonversi gambar dokumen menjadi teks yang dapat diproses lebih lanjut [Firhan Maulana Rusli et al., 2020].

5. Ekstraksi Entitas menggunakan NER
   Setelah teks diekstraksi, dilakukan tokenisasi teks dan pelabelan entitas menggunakan model Named Entity Recognition (NER). Tokenisasi dilakukan menggunakan tokenizer dari library NLP seperti NLTK atau spaCy, dan model NER dilatih menggunakan Large Language Models seperti BERT yang telah disesuaikan dan dilatih ulang pada dataset yang telah diannotasi untuk mengenali entitas seperti nama, lokasi, dan tanggal [Kryeziu and Shehu, 2023]

6. Integrasi Output OCR dan NER
   Output dari OCR kemudian diintegrasikan dengan input model NER. Pengembangan pipeline data dilakukan untuk mengintegrasikan output OCR ke dalam format yang sesuai untuk input model NER, memastikan bahwa teks yang diekstraksi dapat dianalisis dengan tepat oleh model NER. Penyelarasan output OCR sebagai input NER ini bisa dilakukan dalam siklus iterasi hingga ditemukan hasil yang maksimal.

7. Evaluasi dan Optimalisasi
   Sistem kemudian dievaluasi menggunakan metrik seperti akurasi, presisi, recall, dan F1-score untuk menilai efektivitas sistem dalam mengidentifikasi dan mengklasifikasikan entitas dengan benar. Berdasarkan hasil evaluasi, sistem dapat dioptimasi melalui penyesuaian parameter model, peningkatan metode OCR, atau penambahan data latih.

8. Iterasi
   Proses iterasi dilakukan dengan mengulangi siklus pengembangan berdasarkan umpan balik dari evaluasi untuk terus meningkatkan akurasi dan efisiensi sistem

ekstraksi entitas. Setiap tahapan dalam pengembangan sistem ini memastikan bahwa teknologi OCR dan kemampuan pemrosesan bahasa alami dari LLM dimanfaatkan secara maksimal untuk meningkatkan efektivitas ekstraksi entitas dari dokumen multi-format.

3.3 Pendekatan

Pendekatan yang digunakan dalam penelitian ini yaitu dengan melibatkan integrasi teknologi Optical Character Recognition (OCR) dan Natural Language Processing (NLP), khususnya teknik Named Entity Recognition (NER). OCR digunakan untuk mengonversi teks dari gambar dokumen menjadi teks digital yang dapat diproses lebih lanjut. [Firhan Maulana Rusli et al., 2020] menjelaskan bahwa teknologi OCR memungkinkan ekstraksi teks dari gambar, yang kemudian dapat diolah untuk mendapatkan informasi yang berguna. Teknologi ini sangat penting dalam mengotomatiskan proses pengambilan data dari dokumen yang berformat gambar atau cetak.

Setelah teks berhasil diekstraksi menggunakan OCR, langkah selanjutnya adalah penerapan teknik NER untuk mengidentifikasi dan mengklasifikasikan entitas dalam teks tersebut. NER bertujuan untuk menemukan dan mengategorikan segmen teks ke dalam kategori-kategori tertentu seperti nama orang, lokasi, organisasi, dan lainnya. [Kryeziu and Shehu, 2023] menyoroti penggunaan model NER berbasis Large Language Models (LLMs) seperti BERT, yang telah terbukti efektif dalam mengenali dan mengklasifikasikan entitas dari teks dalam berbagai bahasa dan domain.

Integrasi antara OCR dan NER dalam sistem ekstraksi entitas memungkinkan pengolahan dokumen multi-format secara otomatis dan akurat. OCR menyediakan input teks yang diperlukan, sementara NER memberikan kemampuan analisis semantik untuk mengidentifikasi entitas penting dalam teks. Pendekatan ini tidak hanya meningkatkan efisiensi dalam pengolahan dokumen tetapi juga meningkatkan akurasi ekstraksi informasi yang relevan dari dokumen tersebut.

Bibliografi

[A. Ceniza et al., 2018] A. Ceniza, Tom Kalvin B. Archival, and Kate V. Bongo (2018). Mobile Application for Recognizing Text in Degraded Document Images Using Optical Character Recognition with Adaptive Document Image Binarization.

[Akpinar et al., 2018] Akpinar, M. Y., Emeklıgıl, E., and Arslan, S. (2018). Extracting table data from images using optical character recognition text. In *2018 26th Signal Processing and Communications Applications Conference (SIU)*. IEEE.

[Anurag Tiwari, 2021] Anurag Tiwari (2021). Data Extraction from Images through OCR. *International Journal for Research in Applied Science and Engineering Technology*.

[Arora et al., 2020] Arora, K., Bist, A. S., Prakash, R., and Chaurasia, S. (2020). Custom OCR for Identity Documents:ocrxnet. *Aptisi Transactions On Technopreneurship (ATT)*, 2(2):112–119.

[Bektemyssova Gulnara and Akhmer Yerassyl, 2022] Bektemyssova Gulnara and Akhmer Yerassyl (2022). Using Image Processing and Optical Character Recognition to Recognise ID cards in the Online Process of Onboarding. *2022 International Conference on Smart Information Systems and Technologies (SIST)*.

[Chandni Kaundilya et al., 2019] Chandni Kaundilya, Diksha Chawla, and Yatin Chopra (2019). Automated Text Extraction from Images using OCR System. *International Conference on Computing for Sustainable Global Development*.

[Devlin et al., 2019] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

[Duc Phan Van Hoai et al., 2021] Duc Phan Van Hoai, Huu-Thanh Duong, and Vinh Truong Hoang (2021). Text recognition for Vietnamese identity card based on deep features network. *Int. J. Document Anal. Recognit.*

[Firhan Maulana Rusli et al., 2020] Firhan Maulana Rusli, Kevin Akbar Adhiguna, and Hendy Irawan (2020). Indonesian ID Card Extractor Using Optical Character Recognition and Natural Language Post-Processing. *International Conference on Information and Communicatiaon Technology.*

[Hoan Tran Viet et al., 2019] Hoan Tran Viet, Quang Hieu Dang, and Tuan-Anh Vu (2019). A Robust End-To-End Information Extraction System for Vietnamese Identity Cards. *National Foundation for Science and Technology Development Conference on Information and Computer Science.*

[Huang et al., 2021] Huang, K., Qi, P., Wang, G., Ma, T., and Huang, J. (2021). Entity and Evidence Guided Document-Level Relation Extraction. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021).* Association for Computational Linguistics.

[Jeklin Harefa et al., 2022] Jeklin Harefa, Alexander, Andry Chowanda, Emir Haikal, Fedrick, and Stendy Antonio Wiranata (2022). Id Card Storage System using Optical Character Recognition (OCR) on Android-based Smartphone. *2022 International Conference on Electrical and Information Technology (IEIT).*

[Karanrat Thammarak et al., 2022] Karanrat Thammarak, Prateep Kongkla, Y. Sirisathitkul, and Sarun Intakosum (2022). Comparative analysis of Tesseract and Google Cloud Vision for Thai vehicle registration certificate. *International Journal of Electrical and Computer Engineering (IJECE).*

[Kreshnik Vukatana, 2022] Kreshnik Vukatana (2022). Ocr and Levenshtein distance as a measure of image quality accuracy for identification documents. *2022 International Conference on Electrical, Computer and Energy Technologies (ICECET).*

[Kryeziu and Shehu, 2023] Kryeziu, L. and Shehu, V. (2023). Bert based named entity recognition for the albanian language. *Interdisciplinary Journal of Research and Development.*

[Monal Deshmukh and S. Maheshwari, 2019] Monal Deshmukh and S. Maheshwari (2019). Free Form Document Based Extraction Using ML.

[Neha Agrawal and Arashdeen Kaur, 2018] Neha Agrawal and Arashdeen Kaur (2018). An Algorithmic Approach for Text Recognition from Printed/Typed Text Images. *Confluence.*

[Perot et al., 2023] Perot, V., Kang, K., Luisier, F., Su, G., Sun, X., Boppana, R. S., Wang, Z., Mu, J., Zhang, H., and Hua, N. (2023). Lmdx: Language Model-based Document Information Extraction and Localization.

[Rao et al., 2019] Rao, D. T. K., Chowdary, K. Y., Chowdary, I. K., Kumar, K. P., and Ramesh, C. (2019). Optical Character Recognition from Printed Text Images. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, pages 597–604.

[Rifiana Arief et al., 2018] Rifiana Arief, A. Mutiara, T. M. Kusuma, and Hustinawaty (2018). Automated Extraction of Large Scale Scanned Document Images using Google Vision OCR in Apache Hadoop Environment.

[S. Surana et al., 2022] S. Surana, Komal Pathak, Mehul Gagnani, Vidhan Shrivastava, Mahesh T R, and Sindhu Madhuri G (2022). Text Extraction and Detection from Images using Machine Learning Techniques: A Research Review. *2022 International Conference on Electronics and Renewable Systems (ICEARS)*.

[Shi et al., 2019] Shi, M., Huang, J., and Li, C. (2019). Entity Relationship Extraction Based on BLSTM Model. In *2019 IEEE/ACIS 18th International Conference on Computer and Information Science (ICIS)*. IEEE.

[Siddharth Salar Et.al, 2021] Siddharth Salar Et.al (2021). Automate Identification and Recognition of Handwritten Text from an Image.

[Sujata Desai et al., 2020] Sujata Desai, Darshana Rajput, and Kiran Patil (2020). An approach for Text Recognition from Document Images. *2020 IEEE Bangalore Humanitarian Technology Conference (B-HTC)*.

[Vedant Kumar et al., 2020] Vedant Kumar, P. Kaware, Pradhuman Singh, Reena Sonkusare, and Siddhant Kumar (2020). Extraction of information from bill receipts using optical character recognition. *2020 International Conference on Smart Electronics and Communication (ICOSEC)*.

[Wang et al., 2017] Wang, C., Chen, W., and Xu, B. (2017). Named entity recognition with gated convolutional neural networks. In *China National Conference on Chinese Computational Linguistics*.

[Wu et al., 2024] Wu, H., Yuan, Y., Mikaelyan, L., Meulemans, A., Liu, X., Hensman, J., and Mitra, B. (2024). Structured Entity Extraction Using Large Language Models.

[Zhang and Yang, 2018] Zhang, Y. and Yang, J. (2018). Chinese ner using lattice lstm. *ArXiv*, abs/1805.02023.