

BAB 1

PENDAHULUAN

Bab ini menguraikan fondasi fundamental yang melatarbelakangi dilaksanakannya penelitian ini. Pembahasan dimulai dari identifikasi masalah operasional pada proses administrasi pendaftaran wisuda sarjana, hingga perumusan urgensi untuk membangun solusi komputasional berbasis rekayasa *Artificial Intelligence* (AI). Untuk memberikan kerangka kerja yang terarah, bab ini juga memaparkan secara lugas mengenai rumusan masalah, batasan ruang lingkup eksperimen, tujuan yang ingin dicapai, serta kontribusi praktis maupun teoretis dari pengembangan arsitektur sistem otomasi validasi dokumen akademik ini.

1.1 Latar Belakang

Penerbitan ijazah sarjana merupakan proses administrasi final yang memiliki implikasi legal jangka panjang. Ijazah bukan sekadar tanda kelulusan, melainkan dokumen negara yang otentisitas dan keakuratan datanya dilindungi oleh undang-undang. Merujuk pada Peraturan Menteri Pendidikan dan Kebudayaan No. 59 Tahun 2014, setiap data yang tercantum dalam ijazah wajib selaras dengan dokumen identitas sah lainnya. Ketidaksesuaian data, seperti perbedaan satu karakter pada nama lengkap atau kesalahan penulisan tempat dan tanggal lahir, dapat memicu pembatalan verifikasi pada sistem ketenagakerjaan nasional maupun internasional.

Dalam ekosistem administrasi pendidikan tinggi, validitas data mahasiswa merupakan aset krusial yang menentukan akurasi pelaporan pangkalan data pendidikan nasional pada saat penerbitan ijazah pendidikan tinggi. Saat ini, institusi pendidikan mengelola ribuan dokumen legal dalam format PDF, salah satu contohnya yaitu seperti ijazah, yang menjadi rujukan utama verifikasi biodata. Secara operasional, proses validasi dokumen-dokumen untuk penerbitan ijazah universitas ini masih dilakukan secara konvensional, di mana staf administrasi Biro Akademik, Kemahasiswaan dan Kerjasama, Universitas Pembangunan Nasional Veteran Jawa Timur melakukan pemeriksaan visual antara dokumen PDF di layar dengan entri data pada Sistem Informasi Akademik (SIAMIK). Proses manual ini

memerlukan konsentrasi tinggi dan waktu yang signifikan, terutama dalam masa pendaftaran kelulusan (yudisium) dan pendaftaran wisuda untuk penerbitan ijazah sarjana yang melibatkan volume dokumen masif.

Permasalahan utama muncul pada inefisiensi dan tingginya risiko human error dalam proses verifikasi dan validasi secara manual. Dokumen PDF hasil pemindaian (scanning) sering kali memiliki variasi kualitas, sudut kemiringan, dan jenis tulisan menyulitkan proses pembacaan cepat. Penggunaan teknologi Optical Character Recognition (OCR) standar sering kali hanya menghasilkan tumpukan teks mentah tanpa struktur (unstructured text). Sistem kesulitan membedakan entitas penting; sebagai contoh, memisahkan antara nama mahasiswa dengan nama pejabat penandatanganan ijazah yang sama-sama muncul dalam dokumen. Tanpa adanya pemahaman kontekstual terhadap teks, validasi otomatis menjadi sulit dicapai secara akurat.

Tantangan terbesar muncul pada dokumen Ijazah SMA, di mana hingga saat ini masih banyak ditemukan ijazah yang menggunakan berbagai macam variasi desain ijazah berbeda-beda tergantung instansi yang menegluarkan ijazah tersebut. Variasi desain ijazah yang sangat beragam, penurunan kualitas kontras pada dokumen lama, hingga penggunaan kaligrafi formal, menciptakan hambatan besar bagi proses validasi manual maupun otomatis. Staf administrasi seringkali mengalami kelelahan kognitif (cognitive fatigue) saat harus membaca ribuan baris dokumen yang sulit diinterpretasikan. Kesalahan interpretasi manusia terhadap biodata mahasiswa ini merupakan titik lemah utama yang sering menyebabkan kesalahan validasi data pada draf ijazah sarjana yang akan dicetak.

Di sisi lain, teknologi Natural Language Processing konvensional berbasis aturan (rule-based) atau Machine Learning klasik tidak lagi memadai untuk menangani variasi teks yang berasal dari Ijazah SMA yang telah dikonversi secara digital. Dibutuhkan pendekatan deep learning yang memiliki kemampuan representasi bahasa yang dalam. Arsitektur Transformer, melalui model IndoBERT, menawarkan solusi mutakhir. (Wilie dkk., 2020) IndoBERT memiliki pemahaman konteks dua arah (bidirectional) yang dilatih pada korpus bahasa Indonesia yang masif, sehingga mampu memahami konteks administratif bahasa Indonesia secara intuitif (Koto dkk., t.t.)

Pada penelitian sebelumnya yang di lakukan oleh Dimas Saputra yang berjudul Klasifikasi Sertifikat Berdasarkan Mata Kuliah Dalam Program Rekognisi Pembelajaran Lampau Berbasis NLP Menggunakan BERT, Hasil evaluasi menunjukkan bahwa konfigurasi bert-base uncased dengan teknik augmentasi Synonym Replacement menghasilkan performa terbaik, dengan akurasi validasi mencapai 95,54% dan F1-score sebesar 0,97. Beberapa penelitian terdahulu telah berupaya menjawab tantangan ini melalui berbagai pendekatan. Aryanto berhasil mengimplementasikan *Deep Learning* menggunakan CNN untuk pengenalan tulisan tangan dengan akurasi tinggi, namun penelitian tersebut terbatas pada pengenalan karakter tanpa pemahaman konteks entitas. (Aryanto dkk., 2023) Di sisi lain, Baharuddin dan Naufal menunjukkan efektivitas *fine-tuning* pada model IndoBERT untuk klasifikasi teks, tetapi belum mengintegrasikannya dengan proses ekstraksi dokumen fisik secara langsung. (Baharuddin & Naufal, 2023). Terdapat kesenjangan (*research gap*) yang nyata di mana penelitian sebelumnya cenderung memisahkan antara tahap ekstraksi citra dan tahap pemahaman bahasa (NLP), sehingga belum tercipta sebuah alur kerja (*pipeline*) yang utuh untuk validasi otomatis biodata dari dokumen PDF yang heterogen.

Penelitian ini hadir untuk menutup kesenjangan tersebut melalui pendekatan pipeline hibrid yang mengintegrasikan keunggulan ekstraksi teks multimodal dengan pemahaman bahasa alami. Kebaruan penelitian ini terletak pada pemanfaatan Large Language Model (LLM) multimodal sebagai instrumen praproses (preprocessing) guna mengekstraksi teks mentah (raw text) secara presisi dari dokumen mixed-text yang rentan terhadap degradasi fisik dan variasi tulisan tangan. Teks mentah tak terstruktur tersebut kemudian ditransmisikan ke dalam lingkungan komputasi lokal untuk diproses oleh model bahasa IndoBERT yang telah melalui tahapan fine-tuning khusus untuk Named Entity Recognition (NER). Melalui integrasi ini, sistem tidak sekadar mendeteksi karakter visual, tetapi mampu secara cerdas memahami makna kontekstual di balik teks untuk memetakan entitas krusial seperti Nama, Tempat Lahir, dan Tanggal Lahir (Koto dkk., t.t.; Wilie dkk., 2020). Pemanfaatan mekanisme Self-Attention pada arsitektur Transformer memungkinkan sistem untuk melepaskan diri dari batasan metode berbasis aturan (rule-based) yang kaku, sehingga menghasilkan inovasi sistem verifikasi akademik

yang tangguh, adaptif terhadap keragaman format Ijazah SMA di Indonesia, serta mampu menekan risiko human error secara signifikan.

1.2 Rumusan Masalah

Berdasarkan identifikasi fenomena di Biro Akademik dan kesenjangan literatur yang telah dibahas sebelumnya, rumusan masalah dalam penelitian ini difokuskan pada integrasi teknologi ekstraksi teks dan pemahaman bahasa alami untuk otomasi administrasi, maka didapatkan rumusan masalah yang disusun dengan focus sebagai berikut:

- a. Bagaimana mengintegrasikan ekstraksi teks mentah berbasis Multimodal Large Language Model (LLM) dengan proses fine-tuning model IndoBERT-NER untuk mengenali entitas biodata pada dokumen Ijazah SMA mixed-text bertuliskan tangan?
- b. Bagaimana mengukur tingkat performa dan keandalan sistem hibrid tersebut dalam memvalidasi hasil ekstraksi entitas biodata untuk pendaftaran wisuda sarjana terhadap database SIAMIK?

1.3 Tujuan Penelitian

Berdasarkan rumusan masalah yang telah dirumuskan sebelumnya, berikut adalah tujuan penelitian yang disusun secara sistematis dan ilmiah untuk menjawab tantangan validasi biodata mahasiswa pada Biro Akademik:

1. Mengembangkan arsitektur pipeline hibrid yang mengintegrasikan Multimodal Large Language Model (sebagai ekstraktor teks mentah) dengan model IndoBERT-NER untuk mengidentifikasi entitas biodata (Nama, Tempat Lahir, dan Tanggal Lahir) secara presisi pada dokumen Ijazah SMA mixed-text bertuliskan tangan.
2. Mengevaluasi tingkat performa model fine-tuning IndoBERT-NER dan keandalan algoritma Fuzzy String Matching dalam memvalidasi hasil ekstraksi biodata terhadap pangkalan data SIAMIK secara otomatis untuk keperluan pendaftaran wisuda sarjana.
3. Merancang purwarupa (prototype) sistem otomasi validasi dokumen pendaftaran wisuda guna meminimalisir risiko kesalahan interpretasi

administratif dan mengurangi potensi kesalahan cetak blangko ijazah sarjana di tingkat perguruan tinggi

1.4 Manfaat Penelitian

Berdasarkan penjelasan latar belakang dan tujuan penelitian yang telah ditetapkan sebelumnya, penelitian ini memiliki beberapa manfaat. Adapun manfaat penelitian ini yaitu

1. Manfaat Akademis

Penelitian ini memberikan kontribusi keilmuan pada pengembangan bidang Natural Language Processing (NLP) di Indonesia, khususnya dalam memecahkan masalah Information Extraction pada dokumen legal berskala lokal yang menantang. Penelitian ini membuktikan efektivitas pendekatan pipeline hybrid yang mengombinasikan Multimodal Large Language Model untuk akuisisi teks dan arsitektur Transformer (IndoBERT-NER) untuk klasifikasi entitas dalam menangani teks non-digital berjenis mixed-text (campuran teks cetak dan tulisan tangan) pada dokumen Ijazah SMA.

2. Manfaat Praktis

- Bagi Institusi Perguruan Tinggi: Mengoptimalkan efisiensi waktu dalam alur kerja verifikasi pangkalan data pendidikan, serta mengurangi biaya operasional dan kerugian materiil yang timbul akibat kesalahan cetak blangko ijazah negara yang bersifat error prone.
- Bagi Staf Administrasi Akademik: Bertindak sebagai asisten cerdas yang meringankan beban kelelahan kognitif (cognitive fatigue) staf saat harus membaca dan menginterpretasikan ribuan dokumen ijazah SMA fisik yang buram atau memiliki variasi tulisan tangan yang sulit dibaca secara manual pada masa pendaftaran wisuda.
- Bagi Mahasiswa: Menjamin bahwa data ijazah sarjana yang diterbitkan memiliki keakuratan identitas (Nama, Tempat Lahir, dan Tanggal Lahir) yang presisi dan tersinkronisasi secara valid dengan riwayat pendidikan sebelumnya, sehingga tidak menimbulkan kendala legalitas di dunia kerja.

1.5 Batasan Masalah

Dalam penelitian ini dibutuhkan batasan masalah agar terdapat batas yang jelas serta lebih terarah. Batasan masalah yang disusun secara spesifik untuk menjaga focus penelitian adalah sebagai berikut:

- a. Jenis Dokumen dan Bahasa: Dataset primer dibatasi pada dokumen digital (format PDF/citra) hasil pemindaian (scan) Ijazah Sekolah Menengah Atas (SMA) yang menggunakan Bahasa Indonesia.
- b. Ruang Lingkup Entitas: Proses ekstraksi, klasifikasi, dan validasi data dibatasi secara spesifik pada tiga entitas biodata utama, yaitu: Nama Lengkap, Tempat Lahir, dan Tanggal Lahir.
- c. Teknologi Ekstraksi Teks (Preprocessing): Proses ekstraksi teks dari citra dokumen (Optical Character Recognition) tidak menggunakan arsitektur model visi lokal, melainkan memanfaatkan layanan Cloud API Multimodal (Gemini 3.1 Flash Lite Flash). Penggunaan API ini dibatasi murni sebagai instrumen praproses akuisisi teks mentah (text dump), tanpa melakukan klasifikasi maupun restrukturisasi data format CoNLL.
- d. Teknologi Pemahaman Bahasa (NLP): Pemodelan downstream task klasifikasi entitas menggunakan model pra-latih IndoBERT-Base-p2 yang melalui proses fine-tuning berbasis Supervised Learning khusus untuk tugas Named Entity Recognition (NER).
- e. Metode Validasi: Proses pencocokan otomatis (auto-validation) antara hasil prediksi model IndoBERT-NER dengan pangkalan data referensi akademik (SIAMIK) dibatasi menggunakan algoritma Fuzzy String Matching (berbasis Levenshtein Distance).
- f. Infrastruktur dan Privasi Data: Pelatihan (fine-tuning) model IndoBERT-NER, anotasi BIO Tagging, serta proses validasi akhir terhadap database universitas dilakukan secara tertutup di lingkungan komputasi lokal (on-premise). Transmisi citra ke layanan Cloud API dibatasi murni untuk tahap preprocessing ekstraksi teks mentah.