Automating Named Entity Recognition for Indonesian Diplomas via Template-Based Synthetic Data Generation

Ali Isra
Doctoral Program in Information Technology, Gunadarma University, Depok, Indonesia
ali.isra.id@gmail.com

Sarifuddin Madenda
Department of Informatics, Gunadarma University, Depok, Indonesia
sarif@staff.gunadarma.ac.id

Ruddy J. Suhatril
Department of Informatics, Gunadarma University, Depok, Indonesia
ruddyjs@staff.gunadarma.ac.id (corresponding author)

Received: 6 December 2025 | Revised: 12 February 2026 | Accepted: 23 February 2026

 DOI: https://doi.org/10.48084/etasr.16738

ABSTRACT

Automated information extraction from administrative documents, such as higher education diplomas, is critical to streamlining public service verification. However, the development of robust Named Entity Recognition (NER) models for this domain is hindered by the scarcity of publicly available datasets and the prohibitively high cost of manual annotation. The current study addresses this gap by proposing a novel automated pipeline for generating domain-specific NER datasets for Indonesian higher education diplomas. Constructed from 36 distinct higher education templates and 3,460 alumni records, the dataset was synthesized using Direct Label Generation, which leveraged the PDDikti higher education database and a template-rendering mechanism. The generated dataset was evaluated by benchmarking two Pre-trained Language Models (PLMs), where the Indonesian Bidirectional Encoder Representations from Transformers (IndoBERT) outperformed the 522M baseline (F1=0.8048), achieving a superior overall F1-score of 0.8213. To validate real-world utility, a pilot evaluation on held-out, noisy real-world diplomas yielded an F1-score of 0.7491 and a recall of 0.8460. These results confirm that the synthetic generation method effectively bridges the data gap, enabling high-performance extraction even on raw, noisy documents.

Keywords-Named Entity Recognition (NER); dataset generation; higher education diplomas; template rendering; IndoBERT; low-resource domains

I. INTRODUCTION

NER is a significant component of Natural Language Processing (NLP) that identifies and classifies entities such as persons, organizations, locations, and quantities within unstructured text [1, 2]. Extracting structured data from Indonesian diplomas is difficult due to the lack of a standardized national layout, where important attributes, including diploma numbers, student names, and others, which are crucial for data verification, drift across coordinates depending on the issuing institution. In this context, reliable extraction is not merely helpful but strictly necessary to build automated verification systems [3].

A crucial application of this technology lies in the processing of legal and administrative documents, such as higher education diplomas. Verification of higher education credentials in Indonesia constitutes a great administrative burden, often requiring manual validation to ensure authenticity. The automation of this diploma verification process is significant for accelerating public services that rely on these documents as supporting data, such as civil service recruitment and academic admissions. By automating the extraction of key entities, student names, identification numbers, and degree details, institutions can significantly reduce verification time and minimize human error.

However, deploying fine-tuned PLMs for this task depends on the availability of high-quality, domain-specific datasets. Generic NER models often fail to capture the nuanced semantic structures of administrative documents. Higher education

diplomas present unique visual-textual challenges not typically found in general text, including:

- Diverse Formats and Layouts: Diploma attributes often follow complex, non-linear patterns. Text may be arranged in rigid form-like structures, narrative paragraphs, or a combination of both. Recent studies on domain-specific NER, such as in legal corpora, indicate that complex entity structures and intricate boundaries require robust annotation strategies [4].
- Terminological Variations: There is a high degree of inconsistency in how academic terms are represented. For instance, the same study level might be written as "D3," "Diploma 3," or "Ahli Madya" across different institutions. Similarly, metadata structures, such as serial numbers and faculty names, follow specific naming conventions that are not accounted for in generic datasets.
- Entity Disambiguation: Distinguishing between entities with similar surface forms is particularly difficult in this domain. For example, a city name appearing on a diploma could function as a "Place of Birth" or the "Institution Address," requiring the model to rely on local context for accurate classification.

A dedicated dataset for the specific domain of Indonesian higher education diplomas remains unavailable, despite these pressing needs. This scarcity is partly due to the domain's "low-resource" nature, where labeled data are rare and difficult to obtain due to privacy concerns. Furthermore, developing such a dataset using traditional manual annotation methods is time-consuming and expensive, requiring domain experts to label thousands of documents to achieve sufficient variety [5].

This study introduces a novel automated annotation pipeline that leverages the PDDikti database to populate diverse institutional templates. By synthetically generating labeled data, this approach circumvents the prohibitive costs of manual annotation while capturing the structural variance of real-world diplomas. The necessity for such structural precision is supported by recent findings in domain-specific NER (e.g., Arabic Biomedical), which demonstrate that complex, consecutive entities require tailored annotation schemes (such as IOBES or BIES) to be effectively recognized by transformer models [6]. The primary contributions of this research are:

- The development of an automated pipeline for generating and annotating synthetic NER datasets tailored to Indonesian higher education diplomas, addressing data scarcity in this low-resource domain.
- The creation of a high-quality, balanced dataset covering critical entities (e.g., institution identity, graduate data) compliant with national government standards for higher education data.
- A comprehensive evaluation demonstrating that fine-tuning IndoBERT on this dataset achieves an F1-score of 0.8213, significantly outperforming baselines trained on related legal domains and proving the efficacy of synthetic data for complex administrative NER tasks.

II. RELATED WORKS

A. NER Implementations and Evolution

The development of Deep Learning (DL) has revolutionized NER, significantly improving its ability to handle complex linguistic structures compared to traditional rule-based and dictionary-based methods [1, 2]. Early DL approaches, such as Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, enabled automatic learning of features from raw text, achieving efficient performance on general benchmarks [3]. However, transformer-based architectures have altered this landscape by leveraging transfer learning to exploit pre-trained language representations, enabling effective fine-tuning on specific domains with limited data. Although these models offer robust baselines, recent studies indicate that their efficacy drops significantly when applied to specialized domains without targeted fine-tuning on domain-specific corpora.

B. Domain-Specific Challenges and Annotation Schemes

Domain-specific datasets are a prerequisite for achieving state-of-the-art model performance. General NER models tend to underperform when identifying named entities for specific demographic groups or highly specialized domains, due to limited coverage in the training data [7]. Successful domain adaptation has been observed across various fields; for instance, models fine-tuned on chemical texts achieved F1-scores of 0.99 in high-temperature steel research [8], and tools like NERSE demonstrated remarkable performance in the software engineering domain [9].

This necessity is equally evident in the legal and administrative domains, which share formal characteristics with higher education diplomas. GerPS-NER [10] was developed for the German public administration to enable precise analysis of legal norms, while in [11], substantial improvements were achieved in Indian legal text processing by constructing the NER-IPL dataset. The annotation strategy plays a decisive role in model performance. In [4], it was argued that although basic schemes such as Inside, Outside (IO) work for news text, complex administrative entities require more granular boundary detection. Furthermore, in [6], it was emphasized that, in specialized fields, the choice of the annotation scheme, such as IO, Begin, Inside, Outside (BIO), or Inside, Outside, Begin, End, Single (IOBES), directly affects the model's ability to recognize consecutive multi-token entities. These studies demonstrate that generic models fail to capture the fine semantic structures of administrative documents, underscoring the need for specialized corpora with robust tagging strategies.

C. Synthetic and Semi-Automated Dataset Generation

The scarcity of labeled data in low-resource domains has driven research towards synthetic generation and semi-automated annotation strategies. Various approaches have been explored to reduce the prohibitive cost of manual labor. Some researchers have used interactive annotation tools, such as web-based interfaces that incorporate BioBERT, to enable real-time refinement of annotations [12]. Others have integrated knowledge graphs, such as Wikidata and GPT-3, to produce massive synthetic datasets containing multi-categorized entities

[13]. Template-based and semi-automated frameworks have proven particularly effective for structured documents. In [14], it was shown that combining generative models (e.g., GPT NeoX) with simple markup languages produces high-quality annotated sentences with minimal seed data. Similarly, in [15], voting schemes were implemented for Russian literature to reduce manual labor by 94%. This research uses a template-rendering mechanism, building on existing semi-automated approaches to synthesize realistic diploma documents. This method ensures accurate automatic annotation without the bias and noise often introduced by purely generative models, while directly addressing the need for rigorous annotation schemes highlighted in [4, 6].

III. PROPOSED METHODOLOGY

The present study introduces an end-to-end pipeline for generating synthetic, domain-specific NER datasets. The framework comprises four primary stages: (i) Template construction from diploma samples, (ii) dynamic data injection using the PDDikti database, (iii) automated annotation using Direct Label Generation, and (iv) dataset finalization via Stratified Group Splitting.

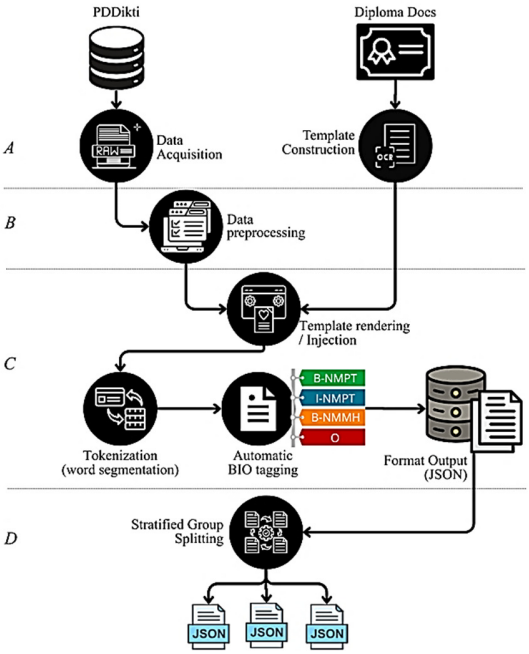

Fig. 1. The proposed synthetic dataset generation pipeline.

A. Acquisition of Raw Entities and Template Extraction

Diploma samples were collected from various Indonesian higher education institutions. A structural analysis of these documents reveals three primary layout patterns: Form-based (key-value-pairs), Paragraph-based (narrative sentences), and Combination layouts. The data comprise 36 diploma templates from various higher education institutions, with 3460 alumni data sourced from the Higher Education Database (PDDikti), as shown in Table I. These data represent the diversity of formats and layouts of higher education diplomas in Indonesia.

TABLE I. LIST ENTITIES AND DATA TEMPLATE

Layout pattern	Diploma samples	Alumni data
Form-based	9	900
Paragraph-based	6	600
Combination	21	1960

Tesseract Optical Character Recognition (OCR) was applied to convert diploma images into text, facilitating the creation of reusable templates. Tesseract OCR can handle various letter types in different sizes and fonts [16]. The choice of Tesseract OCR further offers cost savings compared to cloud-based services like Amazon Textract, despite its superior handling of complex layouts [17]. Before running the OCR engine, Tesseract OCR applied basic image preprocessing steps, such as binarization and slight deskewing, to minimize common OCR errors caused by document misalignment. Specifically, Page Segmentation Mode (PSM) 4 in Tesseract was used, assuming a single-column layout with variable text sizes. This mode was selected because diploma layouts often intersperse large header fonts with smaller body text, causing standard column detection to fail.

Despite preprocessing, the raw OCR output retained artifacts. To quantify the extraction quality, a segmented evaluation was performed to distinguish between machine error and annotation requirements. On static text regions (e.g., labels, institution names), the raw Tesseract output achieved a remarkably low Character Error Rate (CER) of 4.33%, indicating that the OCR engine successfully captured the document's structural skeleton. However, the overall divergence between the raw output and the final templates was 73.95%. This substantial gap quantifies the 'Semantic Abstraction Cost'—the extensive manual effort required not to fix OCR errors, but to systematically replace specific Personal Identifiable Information (PII) with standardized generative placeholders (e.g., converting 'Budi Santoso' to _nama_). To ensure zero-defect generation seeds, a strict Human-in-the-Loop process is enforced, where all of the 36 extracted templates undergo manual curation. The refined text is then used to build a template. Dynamic data are marked as placeholders, later replaced with actual data from PDDikti. Placeholders are made unique by adding underscores at the start and end of each name (e.g., _nama_, _nmpt_) to avoid replacement errors. Several placeholders are designated as entities, and the selection of these entities is governed by Regulation Number 50 of 2024 of the Minister of Education, Culture, Research, and Technology [18]. Table II lists the primary entities and placeholders identified.

B. Data Preprocessing

The PDDikti data to be used in producing the dataset first undergo an imputation process to fill in empty data, and adjustments to the data format are made to match the formatting of the diploma samples used. Furthermore, protecting alumni's personal data is crucial while maintaining the original data representation. The data undergo preprocessing that includes:

- Privacy Protection: Sensitive attributes (e.g., birth dates) are randomized to preserve the overall data distribution while preventing the disclosure of individual records. A ±2-month

range is selected for date randomization. This specific window is small enough to maintain realistic temporal sequence patterns (e.g., correlations between birth year and graduation year) but large enough to effectively mask the original data, thereby preserving distributional integrity.

- Data Imputation: Missing fields (e.g., institutional or study program accreditation data) are imputed based on relevant institutional data to ensure the dataset's completeness. Data imputation is based on relevant alum data with complete accreditation information.

TABLE II. LIST ENTITIES AND DATA TEMPLATE

PDDikti data	Placeholder	Entity
Diploma number	_nina_	NINA
Student name	_nama_	NMMH
Name of college	_pt_	NMPT
Student ID number	_nim_	NIMH
Population identification number	_nik_	NIKM
Place of birth	_tmpt_lahir_	TPLH
Date of birth	_tgl_lahir _	TGLH
Level of study	_jenjang_	JJNG
Study program	_prodi_	PROD
Graduation date	_tgl_yudisium_	TGYD
Degree obtained	_gelar_panjang_	-
Degree abbreviation	_gelar_singkat_	-
Faculty	_fak_	-
First semester of college	_smt_masuk_	-
Entry into the academic year	_thn_ajaran_masuk_	-
Year of starting college	_thn_masuk_	-
Institutional accreditation decree number	_sk_akredpt_	-
Institutional accreditation rating	_peringkat_akredpt_	-
Study program accreditation decree number	_sk_akredprodi_	-
Study program accreditation rating	_peringkat_akredprodi_	-

C. Data Injection Mechanism with Automatic BIO Tagging

The novelty of this method is the Direct Label Generation Pipeline, a function $f: \mathcal{P} \longrightarrow (t, y)$ that automatically generates the token sequence (t) and the corresponding label vector (y) from the rendered paragraph $(\mathcal{P})$ without manual intervention. This approach leverages the known source of the injected data for annotation. In this implementation, the text was rendered cleanly to establish a theoretical upper-bound performance for the architecture, intentionally minimizing visual noise to focus on structural learning.

1) Tokenization and Initiation

Tokenization in this stage differs from BERT Tokenization (sub-word) that occurs later during training. In the specific stage, tokenization breaks sentences into individual words and punctuation, allowing them to be tagged with BIO labels. First, the rendered diploma text $\mathcal{P}$ is tokenized using a regex-based function $tok(\cdot)$ to produce an ordered token list $t = (t_1, \dots, t_n)$. The label vector $y = (y_1, \dots, y_n)$ is initialized by assigning the "Outside" tag (O) to all tokens:

$$t = tok(\mathcal{P}) \quad (1)$$

$$y_i \leftarrow idx(O)\ \forall i \in \{1, \dots, n\} \quad (2)$$

2) Automatic B-I Tag Assignment

For each injected entity $e_k \in \mathcal{E}$, its tokenized form w_k is utilized to locate its exact span within the token sequence t. The system iteratively searches for a match and assigns the corresponding Begin (b_k) and Inside (i_k) tags. The assignment logic is formalized as follows. If a match is found at index i for the tokenized entity w_k:

$$y_i \leftarrow b_k \text{ and } y_{i+j-1} \leftarrow i_k \text{ for } j = 2, \dots, m_k \quad (3)$$

This process generates a fully annotated dataset in JSON format, structured to retain essential metadata, including the *templates* object (containing *name* and *format* fields), along with the *tokens* and *ner_tags*.

D. Stratified Group Splitting Strategy for Partitioning the Final Dataset

The diversity and balance of data in each dataset significantly support the model's generalization process and mitigate the potential for overfitting [19, 20]. Following data generation, Stratified Group Splitting was implemented to ensure model robustness against unseen layouts and prevent data leakage.

Group $(\mathcal{G})$ is defined as the unique template ID (higher education institution), and Stratification Key $(\mathcal{S})$ as the layout format (*form*, *paragraph*, *combination*). The division must meet two key criteria:

- Group Integrity: Each template ID must only appear in one dataset split to prevent the model from memorizing specific template layouts. If G_X is the set of groups in split X:

$$G_{\text{train}} \cap G_{\text{val}} = G_{\text{train}} \cap G_{\text{test}} = G_{\text{val}} \cap G_{\text{test}} = \emptyset \quad (4)$$

- Stratification Balance: The proportion of the layout format $s_k \in \mathcal{S}$ in a split X must approximate the global proportion $\rho(s_k)$:

$$\text{Prop}(s_k|\mathcal{D}_X) \approx \rho(s_k) = \frac{|\{d \in \mathcal{D}: \text{format}(d) = s_k\}|}{|\mathcal{D}|} \quad (5)$$

This method ensures that the model is rigorously evaluated on template structures that it has never encountered during training. Unlike standard random splitting, which allows identical layout structures to appear in both training and testing sets (leading to inflated performance via memorization), this stratified approach strictly forces the model to learn semantic context rather than overfitting to specific spatial patterns. The Stratified Group Splitting strategy partitioned the dataset into a 70:15:15 ratio for training, validation, and testing, which encompassed 36 distinct diploma templates. Strict adherence to group integrity ensures that no single template ID appears in multiple subsets, guaranteeing that the validation and testing phases evaluate only unseen institutional layouts. The resulting distribution confirms that every entity category retains a comprehensive representation of all three layout formats, Combination, Form, and Paragraph, despite variations in total counts. Figure 2 illustrates the balanced composition and layout diversity of the training dataset.

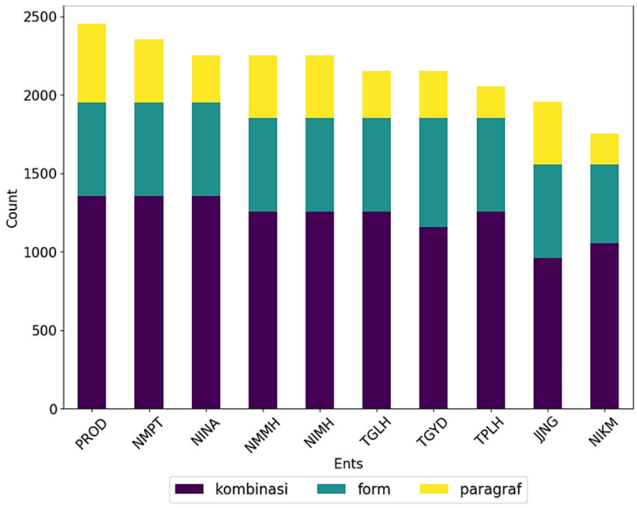

Fig. 2. Entity distribution based on template format in the training dataset.

E. Ethical Considerations and Data Availability

Data access was granted by the Center for Data and Information (Pusdatin) Kemdiktisaintek under the official permit number 168/DST/F1/TI.03.00/2026, in strict compliance with Indonesia's Personal Data Protection Law (UU PDP No. 27 of 2022) [21]. To safeguard privacy, the training dataset containing actual PDDikti records remains confidential and is not publicly released; only the generation pipeline and dummy templates are made available. Furthermore, internal date randomization (±2 months) was applied to prevent overfitting to exact birthdates while preserving Temporal Semantic Consistency, ensuring that logical correlations between cohort years and graduation dates remain intact for effective model training.

IV. EXPERIMENTAL EVALUATION

A. Experimental Setup and Model Performance

The dataset presented in this study was evaluated by fine-tuning two distinct PLMs for NER tasks: IndoBERT base uncased (Indolem) and Indonesian BERT base model 552M (cahya). Both are available in the HuggingFace LLM repository. These models support fine-tuning for NER tasks in the Indonesian college diploma domain because they were trained on Indonesian-language data. To ensure metric integrity across these PLMs, a 'First-Token-Only' alignment strategy was implemented, where labels were assigned strictly to the initial subword of a tokenized term. Subsequent subwords were masked (index -100) during evaluation to prevent score inflation from trivial suffix prediction. Additionally, strict entity-level (span-based) scoring was employed, where only exact matches of both the boundary and the label were counted as correct. Partial matches were explicitly penalized. Micro-F1 (primary) addresses class imbalance, and Macro-F1 per-class consistency.

To ensure full reproducibility, Table III details the exact specifications of the software, hardware, and hyperparameters used for both models. A fixed random seed was used to guarantee deterministic results, and a fixed-epoch training strategy was used based on preliminary convergence analysis.

TABLE III. EXPERIMENTAL DETAILS AND COMPUTATIONAL COST

Component	Specification/value
Dataset splits	Total samples (100%) : 3,460 documents Training set (70%) : 2,422 documents Validation set (15%) : 519 documents Test set (15%) : 519 documents
Hardware	Intel Core i5 (2.4 GHz quad-core) CPU, 8 GB RAM
Software	Python 3, transformers
OCR engine	Tesseract OCR v5.2.0 with config --oem 3 --psm 4
Pre-trained model (HuggingFace)	indolem/indobert-base-uncased cahya/bert-base-indonesian-522M
Tokenizer	Corresponding WordPiece Tokenizer for each PLM
Optimizer	AdamW (learning rate:2e-5, weight decay:0.01)
Training configuration	Batch size: 8, epochs: 3 (fixed, no early stopping), max seq: 512 tokens
Max sequence length	Dynamic (no truncation applied)
Random seed	42
Total training steps	849 Steps (based on actual logs)
Training duration	~2.0 h
Inference latency	~0.15 s / document

Class-wise F1-scores alongside the macro-average metrics are reported to address the class imbalance inherent in the dataset. This granular breakdown allows for a precise assessment of the model's performance on minority entities, ensuring that high overall accuracy does not mask underperformance in specific, less frequent classes.

TABLE IV. EVALUATION RESULTS

Pre-trained model	Entity	Precision	Recall	F1-Score
IndoBERT base case	Ents (micro)	0.8056	0.8376	0.8213
	Ents (macro-avg)	0.8358	0.8836	0.8531
	NINA	1.0000	0.9901	0.9950
	NMPT	0.7357	0.9833	0.8417
	NMMH	0.8542	0.8350	0.8445
	NIMH	1.0000	1.0000	1.0000
	TPLH	0.7063	0.8818	0.7844
	TGLH	0.9967	1.0000	0.9983
	NIKM	1.0000	1.0000	1.0000
	JJNG	0.6036	0.6000	0.6018
	PROD	0.7548	0.5558	0.6402
	TGYD	0.7071	0.9900	0.8250
Indonesian BERT base model 522M	Ents (micro)	0.7449	0.8752	0.8048
	Ents (macro-avg)	0.7548	0.8647	0.8015
	NINA	0.5617	0.8144	0.6648
	NMPT	0.6640	0.8333	0.7391
	NMMH	0.9662	1.0000	0.9828
	NIMH	0.9520	0.8600	0.9037
	TPLH	0.9852	0.9983	0.9917
	TGLH	0.8694	0.9983	0.9294
	NIKM	0.9106	0.9675	0.9382
	JJNG	0.4571	0.6933	0.5510
	PROD	0.5691	0.7417	0.6440
	TGYD	0.6128	0.7400	0.6704

Table IV demonstrates that IndoBERT achieved superior performance, with an overall micro F1-score of 0.8213 and a macro-average of 0.8531, consistently outperforming the distinct baseline, Indonesian BERT 522M (micro F1=0.8048; macro F1=0.8015). This cross-model consistency validates the robustness of the synthetic pipeline across architectures. IndoBERT attained flawless recognition (F1=1.0000) for critical entities such as NIMH (Student ID) and NIKM

(Population ID), while maintaining high precision for NINA (0.9950). These results confirm the precision of the Direct Label Generation process. Furthermore, the high scores on unseen templates validate the Stratified Group Splitting strategy, proving that the model learned structural patterns rather than memorizing documents. The performance advantage of IndoBERT over the general-purpose Indonesian BERT is attributed to its pre-training corpus being better aligned with the formal administrative diction of diplomas.

B. Performance on Real-World Data (Pilot Study)

To assess generalization beyond synthetic artifacts, the model was evaluated on a held-out set of 39 real-world diplomas (spanning 35 distinct templates). Unlike the training data, these samples contain raw OCR noise and underwent no manual text cleaning. The model achieved an overall F1-score of 0.7491 and an overall recall of 0.8460 on this noisy set. Although slightly lower than the synthetic benchmark (0.8213) due to OCR artifacts, this strong performance confirms the model's Sim-to-Real transfer capability. The high recall indicates that the model effectively retrieves entities even in the presence of real-world noise, validating the synthetic dataset as a robust training ground for practical applications.

C. Error Analysis and Limitations

The analysis reveals challenges in both synthetic and real-world scenarios. In the synthetic domain, performance disparities highlight a significant linguistic bottleneck. Although numeric identifiers such as NIKM and NIMH achieved flawless recognition (F1=1.00), extracting the Study Program (PROD) proved significantly more challenging (F1=0.64). This drop is attributed to contextual volatility, where PROD is frequently flanked by dynamic placeholders without stable static anchors, and lexical ambiguity, where terms like 'Teknik' appear in both Faculty and Program names, creating semantic confusion.

Furthermore, the pilot study on real-world data exposed the impact of environmental noise. A Sim-to-Real performance gap was observed, where the model maintained a high recall of 0.8460, proving that it can effectively locate entities, but suffered a drop in precision to 0.6722. This indicates that raw OCR artifacts (e.g., typos, misaligned characters) generate false positives that the model, trained on clean synthetic text, struggles to filter out.

Consequently, this study acknowledges limitations regarding data utility constraints, as privacy concerns limit the publication of the PDDikti raw dataset. Moreover, the observed precision gap in real-world scans confirms the necessity of OCR-noise simulation. Future work will focus on integrating noise injection into the training pipeline to enhance robustness against low-quality inputs.

V. CONCLUSION

This study overcomes the scarcity of Indonesian diploma datasets through a privacy-preserving synthetic generation pipeline. Quantitative analysis validated the curation effort, revealing a 73.95% semantic transformation rate despite a low raw Tesseract Optical Character Recognition (OCR) error of 4.33%. Benchmarking results showed that IndoBERT (F1=0.8213) outperformed the Indonesian Bidirectional Encoder Representations from Transformers (BERT) baseline (0.8048) and achieved flawless recognition for static identifiers. A pilot evaluation on noisy, real-world diplomas confirmed robust Sim-to-Real transferability (F1=0.7491).

Future work will prioritize the integration of OCR-noise simulation and multimodal capabilities (e.g., LayoutLM) to bridge the observed precision gap in real-world scans. Although traditional lightweight models (e.g., BiLSTM or CRF) offer lower computational costs, this study prioritized Transformer-based architectures to handle the significant layout heterogeneity across institutions. To further optimize for edge deployment, future research will also explore knowledge distillation techniques (e.g., DistilBERT) to reduce model size without compromising the semantic generalization capabilities required for variable document structures. Finally, by successfully bypassing strict privacy constraints, this research enables transitioning Indonesian diploma verification into an instantaneous, fraud-resistant digital ecosystem.

DATA AND CODE AVAILABILITY STATEMENT

To facilitate reproducibility while complying with strict data privacy regulations (UU PDP No. 27 of 2022), the anonymized synthetic diploma templates and dataset samples used in this study have been deposited in [22]. However, the source code regarding the core generation and training pipeline is considered proprietary intellectual property. It is available from the corresponding author upon reasonable request for non-commercial research purposes. The raw real-world diploma dataset remains confidential to protect the privacy of the data subjects.

REFERENCES

[1] G. Talukdar, P. P. Borah, and A. Baruah, "Assamese Named Entity Recognition System Using Naive Bayes Classifier," in *Advances in Computing and Data Sciences*, 2018, pp. 35–43, https://doi.org/10.1007/978-981-13-1810-8_4.

[2] A. K. Jumani, M. A. Memon, F. H. Khoso, A. A. Sanjrani, and S. Soomro, "Named Entity Recognition System for Sindhi Language," in *Emerging Technologies in Computing*, vol. 200, M. H. Miraz, P. Excell, A. Ware, S. Soomro, and M. Ali, Eds. Springer International Publishing, 2018, pp. 237–246.

[3] S. Gowr. P and Kumar. N, "Named Entity Recognition for Protecting Sensitive Data using Hybrid CNN," in *2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT)*, Jan. 2023, pp. 1235–1242, https://doi.org/10.1109/ICSSIT55814.2023.10061119.

[4] T. El Moussaoui, C. Loqman, and J. Boumhidi, "Exploring the Impact of Annotation Schemes on Arabic Named Entity Recognition across General and Specific Domains," *Engineering, Technology & Applied Science Research*, vol. 15, no. 2, pp. 21918–21924, Apr. 2025, https://doi.org/10.48084/etasr.10205.

[5] Y. Chen *et al.*, "Prompt robust large language model for Chinese medical named entity recognition," *Information Processing & Management*, vol. 62, no. 5, Sept. 2025, Art. no. 104189, https://doi.org/10.1016/j.ipm.2025.104189.

[6] I. Ait Talghalit, H. Alami, and S. O. El Alaoui, "Exploring Different Annotation Schemes for Single and Consecutive Named Entity Recognition in the Arabic Biomedical Domain using Transformer Models and Contextual Semantic Embeddings," *Engineering, Technology & Applied Science Research*, vol. 15, no. 2, pp. 21854–21860, Apr. 2025, https://doi.org/10.48084/etasr.10019.

[7] N. Loukachevitch *et al.*, "NEREL-BIO: a dataset of biomedical abstracts annotated with nested named entities," *Bioinformatics*, vol. 39, no. 4, Apr. 2023, Art. no. btad161, https://doi.org/10.1093/bioinformatics/btad161.

[8] M. S. U. Miah, J. Sulaiman, T. B. Sarwar, I. U. Ferdous, S. S. Islam, and Md. S. Haque, "Target and Precursor Named Entities Recognition from Scientific Texts of High-Temperature Steel Using Deep Neural Network," in *Database and Expert Systems Applications*, 2023, pp. 203–208, https://doi.org/10.1007/978-3-031-39821-6_16.

[9] M. V. P. Reddy, P. V. R. D. Prasad, M. Chikkamath, and S. Mandadi, "NERSE: Named Entity Recognition in Software Engineering as a Service," in *Service Research and Innovation*, 2019, pp. 65–80, https://doi.org/10.1007/978-3-030-32242-7_6.

[10] L. Feddoul, "GerPS-NER: A Dataset for Named Entity Recognition to Support Public Service Process Creation in Germany," presented at the Second International Workshop on Semantic Technologies and Deep Learning Models for Scientific, Technical and Legal Data SemTech4STLD, May 2024.

[11] S. Jain and P. Harde, "NER-IPL: Indian Legal Prediction Dataset for Named Entity Recognition," in *Business Analytics and Decision Making in Practice*, 2024, pp. 41–50, https://doi.org/10.1007/978-3-031-61589-4_4.

[12] Y. J. Park, M. Lee, G. J. Yang, S. J. Patrk, and C. B. Sohn, "Biomedical Text NER Tagging Tool with Web Interface for Generating BERT-Based Fine-Tuning Dataset," *Applied Sciences*, vol. 12, no. 23, Nov. 2022, Art. no. 12012, https://doi.org/10.3390/app122312012.

[13] A. Belbekri, W. Bouarroudj, F. Benchikha, and Z. Boufaida, "Generating Synthetic Training Data for Named Entity Recognition With Large-Scale Models Integrating Wikidata and GPT," presented at the RIF'24: The 13th Conference on Research in computIng at Feminine, May 2024.

[14] J. Frei and F. Kramer, "Annotated dataset creation through large language models for non-english medical NLP," *Journal of Biomedical Informatics*, vol. 145, Sept. 2023, Art. no. 104478, https://doi.org/10.1016/j.jbi.2023.104478.

[15] K. Kassab, N. Teslya, and E. Vozhik, "Automated Dataset-Creation and Evaluation Pipeline for NER in Russian Literary Heritage," *Applied Sciences*, vol. 15, no. 4, Feb. 2025, Art. no. 2072, https://doi.org/10.3390/app15042072.

[16] J. Park, E. Lee, Y. Kim, I. Kang, H. I. Koo, and N. I. Cho, "Multi-Lingual Optical Character Recognition System Using the Reinforcement Learning of Character Segmenter," *IEEE Access*, vol. 8, pp. 174437–174448, 2020, https://doi.org/10.1109/ACCESS.2020.3025769.

[17] M. Modi, A. Shah, D. Kothadiya, and M. Rahevar, "Optical Character Recognition: Comparative Analysis of Tesseract and Textract on Diverse Datasets," in *2024 3rd International Conference on Automation, Computing and Renewable Systems (ICACRS)*, Dec. 2024, pp. 913–919, https://doi.org/10.1109/ICACRS62842.2024.10841500.

[18] Indonesia, Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi, (2024). *Peraturan Menteri Pendidikan, Kebudayaan, Riset dan Teknologi Nomor 50 Tahun 2024*.

[19] L. Li and M. Spratling, "Data Augmentation Alone Can Improve Adversarial Training." arXiv, 2023, https://doi.org/10.48550/ARXIV.2301.09879.

[20] J. Sheikh, F. Farahnakian, F. Farahnakian, L. Zelioli, and J. Heikkonen, "SEDA: Similarity-Enhanced Data Augmentation for Imbalanced Learning," in *Pattern Recognition*, 2025, pp. 32–47, https://doi.org/10.1007/978-3-031-78395-1_3.

[21] Indonesia, Pemerintah Pusat (2022), *Undang-undang (UU) Nomor 27 Tahun 2022 tentang Pelindungan Data Pribadi*.

[22] A. Isra, S. Madenda, and R. J. Suhatril, "Dataset Samples and Templates for: Automating Named Entity Recognition for Indonesian Diplomas via Template-Based Synthetic Data Generation." Zenodo, Jan. 18, 2026, https://doi.org/10.5281/zenodo.18285791.