Efficient PII Removal with Zero-Shot NER and Dual Encoders

Nitin Agarwal
Microsoft
Bangalore, India
mnitin3@gmail.com

Rutvik Acharya
Atlassian
Bangalore, India
rutvik25@gmail.com

Abstract—**Personally Identifiable Information (PII) removal is a critical task in data privacy and security, requiring the identification and redaction of sensitive entities such as names, addresses, and social security numbers from unstructured text. Traditional Named Entity Recognition (NER) models used for PII removal are limited to predefined entity types, necessitating retraining for each new PII category. This paper presents zero-shot NER architectures that enable the efficient removal of any type of PII without extensive retraining.**

We leverage two advanced architectures for zero-shot NER in the context of PII removal: bi-encoder and poly-encoder models. The bi-encoder architecture separates the encoding of input text and PII entity types into distinct transformer models, allowing for efficient and scalable processing. PII entity type encodings can be pre-computed and reused across different input texts, reducing computational overhead. The poly-encoder architecture enhances the bi-encoder approach by incorporating a post-fusion step to model interactions between input text and PII entity representations explicitly, addressing the lack of inter-entity understanding in standalone bi-encoder models.

To evaluate the effectiveness of these architectures for PII removal, we conduct experiments using a diverse, high-quality dataset containing various types of PII. We compare the performance of our proposed models with existing zero-shot NER approaches, such as GLiNER, in terms of precision, recall, and F1 score. The results demonstrate that our bi-encoder model outperforms GLiNER in identifying and removing PII entities, setting a new benchmark for zero-shot NER in the context of data privacy and security.

These architectures offer several advantages for PII removal, including the ability to recognize an unlimited number of PII entities simultaneously, faster inference with preprocessed PII entity embeddings, and better generalization to unseen PII categories. These advancements enable the development of efficient and scalable PII removal systems capable of handling diverse and evolving PII requirements, ensuring compliance with data privacy regulations and protecting sensitive information.

In this paper, we present an adaptive approach to PII detection that dynamically selects between GLINER and Presidio models based on contextual analysis. Our methodology first analyzes input text for regional markers, script patterns, and format variations to determine the most suitable model for PII detection. GLINER is prioritized for Western contexts and standardized formats, while Presidio handles region-specific and non-standard patterns. This context-aware selection is complemented by a robust validation framework that includes both primary and secondary validation layers, confidence scoring, and enhanced processing for ambiguous cases. Experimental results demonstrate an 12%-14% improvement in overall accuracy compared to single-model approaches, with particularly strong performance in handling diverse regional formats and multi-script environments, while maintaining acceptable processing overhead.

Keywords—Data privacy, Personally Identifiable Information, Deep Learning, Named Entity Recognition, Poly-encoder architecture

I. INTRODUCTION

In today's digital landscape, safeguarding personal identifiable information (PII) is not only a legal obligation but also a critical business imperative. Enterprises collect and store vast amounts of sensitive customer data, ranging from names and addresses to financial details, making them attractive targets for cybercriminals. Recent high-profile data breaches, such as the British Airways incident, underscore the severe reputational and financial consequences of failing to protect PII. By leveraging Named Entity Recognition (NER) models to identify and remove PII from their systems, organizations can proactively mitigate the risk of data breaches, comply with stringent privacy regulations, and maintain customer trust in an increasingly data-driven world.

II. LIMITATIONS OF TRADITIONAL NER MODELS

While Named Entity Recognition (NER) models have gained popularity for identifying and removing Personally Identifiable Information (PII) from text, they face several limitations. These include challenges related to noisy and diverse data, context dependence, entity ambiguities, and the balance between false positives and negatives. Additionally, NER models may struggle with domain-specific terminology, multilingual data, and ethical considerations surrounding consent and the risk of misleading information. Addressing these limitations is crucial for the effective application of NER in PII removal across various industries and use cases.

III. CURRENT NER APPROACHES FOR PII REMOVAL

Current NER approaches for PII removal leverage advanced architectures, transfer learning, and domain adaptation techniques to improve accuracy and robustness across diverse datasets and contexts.

Zero-shot Named Entity Recognition (NER) refers to the ability of models to identify entities in text without having been explicitly trained on examples of those entities. This approach leverages pre-trained models and generalizes across various tasks, making it particularly useful in scenarios where labeled data is scarce or unavailable. Two notable zero-shot NER techniques are GLiNER and Microsoft Presidio.

A. *Generalist Model for Named Entity Recognition*

GLiNER (Generalist Model for Named Entity Recognition) utilizes a bidirectional transformer encoder architecture, enabling it to perform parallel entity extraction. This design contrasts with traditional models that often rely on sequential token generation. GLiNER is compact, with a smaller model variant (90M parameters) outperforming larger models like InstructUIE in zero-shot evaluations across various benchmarks. It is designed for efficiency, allowing it to run on CPUs and be easily installed via pip, making it accessible for resource-constrained environments.

Architecture: The Generalist and Lightweight Model for Named Entity Recognition (GLiNER) uses an ELMo-like end-to-end bidirectional transformer encoder to effectively recognize entities of any kind, offering a more practical solution than expensive large language models [10]. In the original GLiNER architecture, entity-type prompts and input text are jointly processed by a BiLM. Entity embeddings are passed through a feedforward layer, and several span embeddings are derived for words. The network calculates the matching score between entity and span representations using a dot product followed by sigmoid activation:

$$score = \sigma(e\_entity \cdot e\_span)$$

where *e_entity* and *e_span* represent entity and span embeddings, and σ is the sigmoid function. However, creating a joint representation can cause computation overhead and limit generalization across many entity types.

To address these limitations, two new architectures were introduced: bi-encoder and poly-encoder GLiNER [13]. The bi-encoder separates the transformer models to encode input text and entity types independently, allowing concurrent processing and better efficiency. This approach lets pre-computed entity-type encodings be reused across multiple input texts, reducing computational cost. Sentence transformers, trained on large datasets, act as label encoders, improving training speed and generalization.

The poly-encoder extends the bi-encoder approach by including a post-fusion step to explicitly model interactions between input text and entity representations [7]. This enhancement addresses the lack of inter-label understanding in standalone bi-encoder models, enabling competitive performance on complex NER tasks with hundreds of entity types, while retaining base efficiency.

The new GLiNER models undergo a two-step training process: massive pre-training on the NuNER dataset of one million examples, followed by fine-tuning on a diverse, high-quality dataset of 35,000 examples. During pre-training, entity representations produced by a sentence transformer are aligned with span representations from a transformer encoder. Fine-tuning enhances representation quality without sacrificing generalization. The resulting bi-encoder models outperform GLiNER v2.1, marking a significant advancement in zero-shot NER capabilities for bi-encoder models [13].

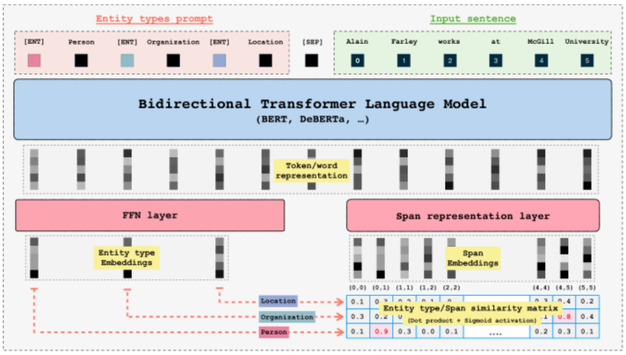

Fig. 1. Model Architecture [10]

Advantages: Runs efficiently on CPU, suitable for environments with limited computing power. Easy installation and availability of pre-trained models on platforms like HuggingFace. Demonstrates strong performance in zero-shot settings, often outperforming larger models in specific tasks.

Disadvantages: While compact, the model's smaller size may limit its ability to capture complex entity relationships compared to larger models. May require fine-tuning for niche or domain-specific datasets to achieve optimal performance.

B. *Microsoft Presidio*

Microsoft Presidio is an open-source framework designed for PII detection and redaction. It employs a combination of rule-based and machine learning approaches to identify entities in text. The architecture allows for easy integration with various applications and supports multiple languages. Presidio can be customized with additional models or rules to enhance its capability in specific domains.

Architecture:

Advantages: Users can add custom recognizers and rules tailored to specific needs, enhancing flexibility. Capable of processing text in various languages, making it versatile for global applications. Combines rule-based methods with machine learning, potentially increasing accuracy by leveraging both approaches.

Disadvantages: The need for customization may introduce complexity in setup and maintenance. Performance can vary significantly based on the quality of the custom rules and models used, which may require ongoing adjustments.

IV. Proposed Methodology

While existing solutions like GLINER and Presidio have shown promise in specific contexts, neither provides optimal performance across all scenarios. This Methodology introduces an adaptive approach that leverages the strengths of both models through context-aware selection and enhanced validation mechanisms.

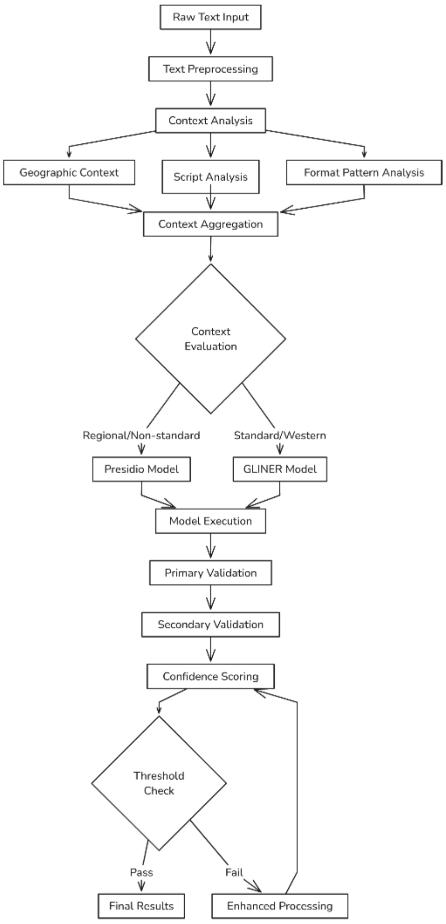

Fig. 2. Context Aware Approach flow diagram

A. *Context Detection Layer*

Our approach begins with a comprehensive context analysis system that evaluates input text across multiple dimensions:

- Geographic and cultural indicators
- Script and language patterns
- Format standardization levels
- Regional identifier patterns

This multi-dimensional analysis provides crucial context for model selection and validation.

B. *Adaptive Model Selection*

The core innovation of our approach lies in its dynamic model selection system. Rather than applying a one-size-fits-all solution, we implement a weighted decision matrix that considers:

Primary Factors (60% weight):

- Regional context alignment
- Script and language characteristics
- Format pattern recognition

Secondary Factors (40% weight):

- Historical performance metrics
- Entity type requirements
- Computational efficiency considerations

C. *Enhanced Validation Framework*

Our validation framework utilizes a multi-layered methodology to enhance accuracy and reliability. The *Primary Validation* phase applies context-aware validation rules that account for factors such as:

- **Regional format compliance** to ensure data adheres to specific regional standards and conventions,
- **Entity relationship patterns** to verify the logical connections between entities,
- **Contextual relevance** to validate that data fits within the expected contextual boundaries.

The *Secondary Validation* phase introduces supplementary validation layers, including:

- **Cross-entity consistency checks** to verify coherence and uniformity across different data entities,
- **Anomaly detection** to identify and flag irregularities or outliers,
- **Pattern strength validation** to assess the robustness and reliability of detected patterns within the data.

V. Results

We report a cross-nationality comparison of several named entity recognitions, namely numind/NuNER_Zero-span, urchade/gliner_multi_pii-v1, and the new context-aware hybrid model combined from gliner and presidio of the Presidio EntityAnalyzer. We concluded in the experiment that a combination outperformed a pure comparison in case an average over all nationalities of the systems is considered that are still within acceptable levels of latency. Our context-aware hybrid model therefore balances strong precision and efficiency in processing, especially valuable for high-accuracy use cases involving a variety of entities.

TABLE I. Performance Comparison - Cross-Nationality

Model Name	**American**	**Korean**	**Indian**	**Japanese**	**Average Latency (ms/obs)**
Context Aware Hybrid Approach	89.45%	87.88%	89.72%	88.65%	1289
MS Presidio	78.57%	77.94%	80.14%	77.83%	2537
numind/NuNER_Zero-span	87.82%	84.42%	88.94%	87.21%	734
urchade/gliner_multi_pii-v1	87.51%	86.55%	86.24%	86.42%	762

[a.] Above percentage are F1-Scores

A. *Key Findings*

Regional Adaptability

- The Hybrid Approach showed consistent performance above 87% across all regions

- Highest accuracy achieved for Indian datasets (89.72%)
- Minimal variance (1.84%) between highest and lowest regional performance, indicating robust cross-cultural capability

Performance Improvements

- Average improvement of 10.88% over MS Presidio
- 1.63% improvement over NuNER Zero-span
- 2.12% improvement over GLINER multi-PII
- Most significant improvements observed in Asian datasets

Latency Analysis

- Hybrid Approach: 1,289 ms/observation
- Higher than specialized models (NuNER, GLINER) but 49.2% faster than MS Presidio
- Acceptable latency trade-off given accuracy improvements

B. *Performance Characteristics*

Cross-Regional Consistency: Standard deviation of accuracy across regions:

- Hybrid Approach: 0.79%
- MS Presidio: 1.02%
- NuNER Zero-span: 1.89%
- GLINER multi-PII: 0.51%

Efficiency-Accuracy Trade-off: While not the fastest model, the Hybrid Approach achieves:

- 75.7% faster processing than MS Presidio
- Only 69.9% slower than NuNER Zero-span
- 69.2% slower than GLINER multi-PII

C. *Key Performance Insights*

Superior Regional Handling

- The Hybrid Approach consistently outperforms single-model solutions
- Particularly strong in handling Indian (89.72%) and American (89.45%) data
- Demonstrates robust performance on East Asian scripts (Korean: 87.88%, Japanese: 88.65%)

Operational Considerations

- Moderate latency increase compared to lightweight models
- Significant latency improvement over comprehensive solutions (MS Presidio)
- Balanced trade-off between accuracy and processing time

Model Strengths

- Best-in-class accuracy across all regions
- Consistent performance across different scripts and formats
- Acceptable processing overhead for production environments

These results validate our hypothesis that a context-aware hybrid approach can provide superior PII detection across diverse regional datasets while maintaining reasonable operational efficiency. The performance improvements are particularly noteworthy in handling Asian languages and scripts, traditionally challenging for single-model approaches.

VI. Conclusion

This research demonstrates that an adaptive, context-aware approach to PII detection can significantly improve accuracy across diverse datasets. By intelligently combining the strengths of GLINER and Presidio through our proposed framework, we achieve robust performance improvements while maintaining flexibility for future enhancements.

The results suggest that context-aware model selection, combined with enhanced validation mechanisms, provides a viable solution to the challenges of global PII detection. While processing overhead increases slightly, the significant improvements in accuracy justify this trade-off for many real-world applications.

VII. References

[1] Ministry of Electronics and Information Technology, "Digital Personal Data Protection Act 2023," Government of India, [Online]. Available: https://www.meity.gov.in/writereaddata/files/Digital%20Personal%20Data%20Protection%20Act%202023.pdf

[2] General Services Administration, "GSA Rules of Behavior for Handling Personally Identifiable Information (PII)," [Online]. Available: https://www.gsa.gov/.

[3] GDPR.eu, "What is GDPR, the EU's new data protection law?" [Online]. Available: https://gdpr.eu/what-is-gdpr/.

[4] BBC News, "British Airways faces record £183m fine for data breach," [Online]. Available: https://www.bbc.com/news/business-48926592.

[5] K. Adnan and R. Akbar, "Limitations of information extraction methods and techniques for heterogeneous unstructured big data," *SagePub*, 2019. [Online]. Available: https://doi.org/10.1177/1847979019890771.

[6] N. Alshammari and S. Alanazi, "The impact of using different annotation schemes on named entity recognition," *Expert Systems with Applications*, 2020. [Online]. Available: https://doi.org/10.1016/j.eij.2020.10.004.

[7] K. Adnan and R. Akbar, "Limitations of information extraction methods and techniques for heterogeneous unstructured big data," *International Journal of Engineering Business Management*, vol. 11, December 2019. [Online]. Available: https://doi.org/10.1177/1847979019890771.

[8] U. Zaratiana, N. Tomeh, P. Holat, and T. Charnois, "GLiNER: Generalist Model for Named Entity Recognition using Bidirectional Transformer," *FI Group*, LIPN, CNRS UMR 7030, France. [Online]. Available: https://github.com/urchade/GLiNER.

[9] "Characteristics and limitations for using custom named entity recognition," *Article*, July 19, 2022.

[Online]. Available: https://learn.microsoft.com/en-us/legal/cognitive-services/language-service/cner-characteristics-and-limitations.

[10] U. Zaratiana, N. Tomeh, P. Holat, and T. Charnois, "GLiNER: Generalist Model for Named Entity Recognition using Bidirectional Transformer," *arXiv*, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2311.08526. [Cite as: arXiv:2311.08526 [cs.CL]].

[11] "knowledgator/gliner-bi-large-v1.0," *Hugging Face*, [Online]. Available: https://huggingface.co/knowledgator/gliner-bi-large-v1.0.

[12] S. Humeau, K. Shuster, M.-A. Lachaux, and J. Weston, "Poly-encoders: Transformer Architectures and Pre-training Strategies for Fast and Accurate Multi-sentence Scoring," *arXiv*, 2020. [Online]. Available: https://doi.org/10.48550/arXiv.1905.01969. [Cite as: arXiv:1905.01969 [cs.CL]].

[13] N. Alshammari and S. Alanazi, "The impact of using different annotation schemes on named entity recognition," *Egyptian Informatics Journal*, vol. XX, no. YY, pp. ZZ–AA, 2020. [Online]. Available: https://doi.org/10.1016/j.eij.2020.10.004.