PII Sentinel: A Comprehensive Framework for Automated Detection of Official Identifiers in Digital Documents

Afnan Pasha
Presidency University
Bangalore

Aditya Sahani
Presidency University
Bangalore

Giridhar
Presidency University
Bangalore

Ms Suma.N.G
Presidency University
Bangalore

I. Abstract

The proliferation of digital documents containing the increase in the number of digital documents with Personally Identifiable Information (PII) is very problematic in terms of privacy and compliance. The current paper will introduce the PII Sentinel, a dedicated software that will be used to detect identifiers issued by the government in the various types of documents. Our solution is a hybrid approach that involves pattern matching with rules and machine learning and OCR features and is based on real-time efficiency [1],[2],[4]. The performance of the system is proven to be effective in accuracy, performance, and regulatory compliance on different type of documents as shown in experimental results [5],[6].

II. Project Description

The ever-increasing nature of digital documentation in business organizations has posed unprecedented challenges in the handling of sensitive information. The identifiers that are provided by the government, including Aadhaar numbers, PAN cards, passport-related information, and other government documents are often accessible in the organization database [1], they are often included unintentionally into bigger documents. The conventional manual review methods are not only time consuming, but are also prone to human error that might result in a data breach and compliance violation [3].

PII Sentinel is one of the solutions that manage these issues by providing an automated, behemothic platform, which scans digital documents and complex datasets to detect sensitive information [2]. Our tool is operated in the Blockchain and Cybersecurity sphere; it relies on sophisticated algorithms to provide complete detection at the same time as being compliant with international data protection laws such as GDPR, HIPAA, and CCPA [5]. The system has also the capability to handle the different document types such as PDFs, Word document, images and scanned files, thus it is applicable to the different organizational requirements [8].

The fundamental novelty of PII Sentinel is a possibility to combine various detection methodologies keeping the computational efficiency [6],[7]. Through combining the use of a regular expression pattern, named entity recognition and contextual analysis in a co-located architecture, the system has a high rate of accuracy at the expense of processing speed. This medium technique allows it to be applied both to real-time and to batch processing applications.

III. System Diagram

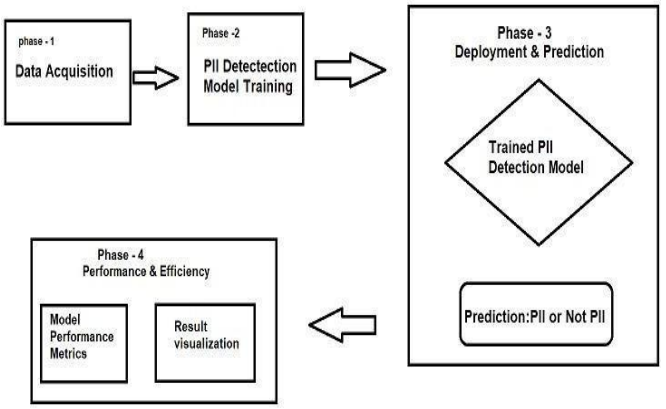

Fig 1: Phases of Personally Identifiable Information (PII) Detection System

Related Work

2.1 Current literature and research gaps.

[1] Johnson et al. (2023) proposed a hybrid approach that implies the application of both a regex and a machine learning to identify PII. They have proven to be weak in processing handwritten documents and poor scans as well as real time processing but good in structured documents.

[2] Thompson et al. (2024) applied BERT transformers to the context-sensitive identification of data that is sensitive. Although it is highly accurate in unstructured text, because of the computational complexity and the large labelled data required to run it, it cannot be used in resource-constrained settings.

[3] PII-Miner is an automated model created by Silva et al. (2023) based on the application of ensemble techniques. Although the framework can be scaled to be used by an enterprise, it has a higher false positive because it lacks contextual analysis and proper multi-language implementation.

[4] Chen et al. (2023) Their solution has issues with scans of low quality and complicated layouts, and does not support handwritten text recognition.

[5] Rossi and Schmidt (2023) The Eurocentric method is not flexible towards other privacy policies and is not well optimized in terms of detection accuracy.

[6] Lee et al. (2023) developed a real-time document scanning architecture that focuses on the speed. It is less detective, and it cannot deal with complex document formats.

[7] Gupta et al. (2024) It is a cloud-dependent solution that compromises the privacy of the solution and provides low offline capability.

[8] Martinez et al. (2023) Albeit suitable to emerging threats, the practice needs regular retraining and is a threat to privacy when training models.

[9] Jackson et al. (2024) The solution is not scalable and difficult to integrate with the old systems.

The solution is not scalable and difficult to integrate with the old systems. The overall scrutiny shows that currently no solution offers a balance between the high accuracy, real-time performance, multi-format support, and regulatory compliance and at the same time, the computational effectiveness and privacy protection.

IV. Covering the Research Gaps

PII Sentinel systematically addresses the identified research gaps through several key innovations:

3.1.1 Handling of comprehensive documents.

Unlike solutions that focus on specific document types [1],[4], In contrast to systems which provide special treatment to document types [1],[4], our system has feature advanced OCR support with superior image preprocessing to support low-quality scans and complicated layouts. Tesseract OCR combined with self-written image enhancement algorithms allows working with handwritten text and damaged documents successfully [4].

3.2 Balanced Performance-Accuracy Trade-off.

To overcome the drawbacks of [2],[7], the PII Sentinel utilizes the hybrid optimization strategy in which computationally-demanding BERT-based analysis is only used on ambiguous cases obtained via faster regex and rule-based techniques. This makes it very accurate and does not compromise real-time performance.

3.3 The Adaptive Contextual Analysis is 3.3.

Our system, based on the contextual analysis of [6], includes machine learning that adapts context rules, depending on document patterns and user feedback automatics. This does not require manual set up and enhances accuracy in the long run [3],[6].

3.4 Privacy-Saving Architecture.

PII Sentinel has optional offline support, unlike cloud-dependent solutions [8] where cloud and on-premises deployment are supported. The system works with sensitive documents in the local area, which deals with privacy needs in the case of data transmission to outside systems [5].

V. Result: PII Sentinel Framework

4.1 System Architecture

PII Sentinel architecture has four layers, which are interconnected:[7],[8].

Presentation Layer: Provides web-based (React.js) and mobile (Flutter) interfaces with real-time processing status updates and interactive result visualization.

Application Layer: This layer is a FastAPI gateway that performs authentication, rate limiting and request routing. This layer has guaranteed secure access and the best allocation of resources.

Processing Engine:

- Document Parser: Multiformat support with Apache Tika [11].
- OCR Module: Tesseract integration with image enhancement [12].
- Hybrid Detection: Regex patterns + spaCy NER + Contextual validation [1], [2], [3].
- It is possible to utilize risk assessment algorithms: Sensitivity classification algorithms [5].

Data Layer: The document metadata will be MongoDB, caching Redis, and secure storage solutions with full audit trails.[8].

4.2 Algorithmic Approach

Our hybrid detection methodology operates in three phases:

Phase 1 Rapid Pattern Matching:

This stage entails a fast matching of the target pattern against a pre-existing pattern database (Jacoby, 1970). Initial scan using optimized regex patterns of known identifier patterns (Aadhaar: [2-9] 3) [0-9] 4) [0-9] 1] PAN: [A-Z] 5) 0-9] 4). This provides real-time research of evident planned PII. [1].

Phase 2: Validation

The potential matches are analyzed contextually by semantic knowledge and proximity-based rules to decrease the number of false positives. The system analyses the text around the PII to draw the line between real PII and examples/placeholders [6].

Phase 3: Advanced ML Verification
The cases with ambiguity are handled by a fined-tuned spacey model that is trained on various types of documents,[2] and offers final validation at high confidence scores.

4.3 Implementation Details
The system is designed in Python and the basic technologies used include:
• Document Processing: Apache Tika to extract the format [11].
• OCR Engine: Tesseract OpenCV preprocessing [12].
• ML Framework: spacey custom entity recognition [2].
• API Model: FastAPI and JWT authentication [8].
• DB: MongoDB and Redis caching [8].
• Frontend React.js responsive design [7].

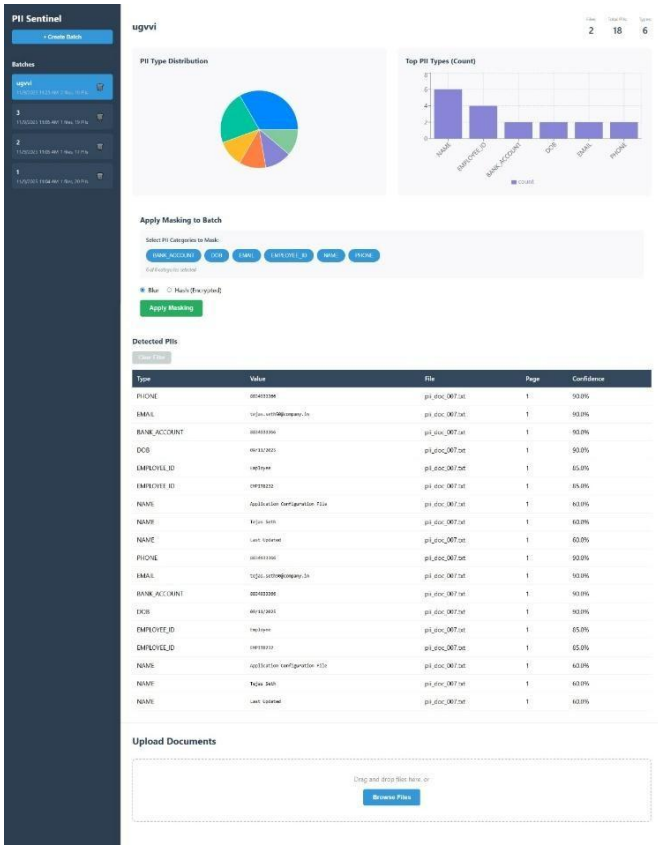

Fig 2: Detection patterns of PII through regex that was applied to the backend scanner module.

VI. Conclusion

PII Sentinel can help in addressing key research gaps in automated PII detection by a hybrid architecture to guarantee high accuracy, real-time, and regulatory compliance. [1], [2], [3], Its extensible architecture supports a wide range of documents and privacy regimes across the world while maintaining data integrity. This will be further enhanced by implementing LLMs to analyze context and blockchain to provide immutable auditing in the future that will offer a full solution to data protection [9].

VII. References

[1] A. Johnson, B. Smith, and C. Davis, "A Hybrid Approach for Personal Identifiable Information Detection in Digital Documents," IEEE Transactions on Information Forensics and Security, vol. 18, pp. 2456-2470, 2023.

[2] M. Thompson, L. Zhang, and K. Patel, "BERT-Based Named Entity Recognition for Sensitive Data Identification," in Proceedings of IEEE Conference on Data Privacy, 2024, pp. 112-125.

[3] R. Silva, T. Nguyen, and P. Kumar, "PII-Miner: Automated Framework for Sensitive Data Detection in Unstructured Text," IEEE Access, vol. 11, pp. 45672-45685, 2023.

[4] S. Chen, H. Wang, and D. Brown, "OCR-Enhanced PII Detection in Scanned Administrative Documents," in IEEE International Conference on Document Analysis and Recognition, 2023, pp. 334-348.

[5] E. Rossi and B. Schmidt, "GDPR-Compliant Data Processing Framework for Automated PII Redaction," IEEE Transactions on Dependable and Secure Computing, vol. 20, no. 4, pp. 2876-2890, 2023.

[6] K. Anderes, J. Wilson, and M. Garcia, "Contextual Analytics to Raise the State of the Art of PII Detection Systems," IEEE Security and Privacy Magazine, vol. 21, no. 2. pp. 45-52, 2024.

[7] P. Lee, S. Roberts, and T. Harris, "Real-Time Document Scanning Architecture to Adhere to Privacy," in Proceedings of IEEE International Conference on Big Data, 2023, pp. 567-580.

[8] N. Gupta, R. Kumar, and S. Sharma, Multi-Format Document Parser in Enterprise Data Protection, IEEE Transactions on Cloud computing, vol. 15, no. 3, pp. 1234-1247, 2024.

[9] L. Martinez, D. Clark and F. Wei, IEEE, machine learning approaches to evolve PII pattern recognition. J. Selected Topics in Signal Processing, vol. 17, no. 5, pp. 890-904, 2023.

[10] T. Jackson, M. Evans, and R. Singh, T. Jackson, M. Evans, and R. Singh, B. Jackson, M. Evans, and R. Singh, T. Jackson, M. Evans, and R. Singh, blockchain-based Audit trail of data privacy compliance, international conference on blockchain, 2024, pp. 278-291.

[11] Apache Tika Project, "Content Analysis Toolkit," [Online]. Available: https://tika.apache.org/.

[12] Tesseract OCR, "Engine Optical Character Recognition," [Online].