10.48047/jocaaa.2019.27.07.09

Cognitive Document Redaction: A Hybrid NER–Regex Pipeline for Sensitive Entity Masking

Partha Sarathi Manuri
Independent Researcher, India

Abstract

The secure handling of sensitive documents, such as contracts and service-level agreements (SLAs), requires robust redaction of Personally Identifiable Information (PII), Non-public Personal Information (NPI), and other confidential data to prevent security breaches. Cognitive document redaction aims to automatically identify and mask sensitive information such as PII, NPI, and financial and location details to mitigate privacy and security risks in unstructured text repositories including contracts and SLAs. Traditional manual redaction is labor-intensive and prone to error. This paper proposes a hybrid pipeline combining pattern-based recognition via regular expressions for high-precision, format-regular entities (e.g., emails, URLs, phone numbers) with statistical Named Entity Recognition (NER) using spaCy for context-dependent entities (e.g., PERSON, ORG, GPE/LOC, MONEY) and extensible custom entity types trained with span-level annotations. Evaluations on adapted subsets of the i2b2/UTHealth de-identification corpus demonstrate superior precision-recall tradeoffs over regex-only and NER-only baselines, with improvements for redaction consistency and content preservation, and strong scalability characteristics suitable for enterprise deployment alongside human-in-the-loop workflows.datatracker.ietf+1

Introduction

Sensitive attributes embedded in operational documents pose compliance and leakage risks; automated redaction reduces manual effort and error rates but must maintain high recall without degrading document utility. Document redaction, the process of obscuring or removing such sensitive information before publication or sharing, is a critical compliance and security task. Prior shared tasks in clinical de-identification highlighted the difficulty of comprehensive PHI detection across varied formats and contexts, motivating hybrid rule-learning approaches to capture both structured patterns and contextual entities. This work targets a practical design using spaCy NER and RFC-informed regexes to redact and replace entities with typed placeholders such as [PHONE], [SSN], [LOCATION], [NAME], and [ORG] for robust downstream sharing.spacy+2

Manual redaction is the traditional solution, but it is notoriously slow, costly, and susceptible to human error, especially when dealing with large document sets. Early automated approaches often relied on simple keyword matching or rule-based systems, which, while fast, struggle with the variability and context-dependency of natural language.projectfusion

To address these limitations, we propose a hybrid model for "Cognitive Document Redaction." This model combines the strengths of two powerful techniques:

1. **Pattern-Based Recognition:** Utilizes regular expressions (regex) to identify and redact entities with a well-defined structure, such as email addresses, phone numbers, and Social Security Numbers (SSNs). This

10.48047/jocaaa.2019.27.07.09

method offers high precision for predictable data formats.dbgroup.eecs.umich

2. **Named Entity Recognition (NER):** Employs a machine learning model to identify entities that require contextual understanding, such as names of individuals (NAME), organizations (ORG), and locations (LOCATION).pmc.ncbi.nlm.nih

Literature Review

The 2014 i2b2/UTHealth shared task provided longitudinal clinical narratives with PHI annotations, showing de-identification systems based on CRFs and hybrids achieving strong F1 yet varying by category and tokenization strategies. Earlier tools like MITRE's MIST used CRF models and noted that some PHI categories are better captured by rules, underscoring complementarity between statistical and pattern-based methods. Contemporary spaCy introduced improved neural NER with subword features and "Bloom" embeddings, enabling robust general-purpose entity extraction and straightforward updating for domain-specific labels, which is well-suited for enterprise redaction pipelines circa 2018.spacy+2

Methodology

Hybrid architecture: The pipeline first applies deterministic regex detectors for pattern-regular entities, then applies spaCy NER for contextual entities, followed by a unification stage that resolves overlaps and assigns canonical placeholders per label taxonomy. Regex module: Patterns are grounded in widely adopted specifications, e.g., email formats guided by RFC 5322 principles for mailbox syntax, pragmatic URL patterns, and locale-aware phone formats, emphasizing maintainable patterns with high precision to minimize false positives. This initial stage targets entities with consistent and predictable formats. We use Python's built-in re library to implement a set of regular expressions for common PII and sensitive data types. This approach is fast and highly accurate for its targets. Examples of regex patterns include:

- **Email Address:** [a-zA-Z0-9._%+-]+@[a-zA-Z0-9.-]+\.[a-zA-Z]{2,}
- **Phone Number (U.S.):** $$?\d{3}$$?[-.\s]?\d{3}[-.\s]?\d{4}
- **Social Security Number:** \d{3}-\d{2}-\d{4}
- **Web URL:** https?:\/\/(www\.)?[-a-zA-Z0-9@:%._\+~#=]{1,256}\.[a-zA-Z0-9()]{1,6}\b([-a-zA-Z0-9()@:%_\+.~#?&//=]*)

When a match is found, the identified text is immediately replaced with its corresponding label, such as [EMAIL], [PHONE], or [SSN]. Statistical NER: spaCy English pipelines extract default labels (PERSON, ORG, GPE/LOC, MONEY/DATE) and can be fine-tuned with span annotations specifying text, start, end, and label for custom entities such as ACCOUNT_NO, POLICY_ID, or CONTRACT_ID, enabling recognition beyond generic ontologies. Conflict resolution and placeholdering: When regex and NER overlap, deterministic precedence is given to the more specific detector (e.g., EMAIL over PERSON), then normalized to placeholders like [EMAIL], [URL], [PHONE], [PERSON], [ORG], [GPE], [MONEY], with optional subtype tags for auditability.

Training and updating: Custom entities are added by updating spaCy's NER with curated examples and negative contexts following the training loop, ensuring stable performance and avoidance of catastrophic interference with out-of-the-box labels.datatracker.ietf+2

10.48047/jocaaa.2019.27.07.09

Datasets

Experiments use a de-identified, institutionally shareable subset adapted from the 2014 i2b2/UTHealth corpus, which contains varied clinical record types, diverse PHI categories, and well-documented annotation guidelines aligned to HIPAA classes, enabling realistic evaluation of person names, organizations, locations, dates, and contact identifiers. The corpus history details category distributions and prior benchmark ranges, providing context for measuring improvements of hybrid methods over single-technique baselines within a reproducible redaction setup. Surrogate PHI in the corpus supports evaluating placeholder fidelity and the preservation of non-sensitive content semantics after redaction.i2b2+2

Implementation Details

spaCy pipeline: The system uses the English NER component with updated weights from in-domain examples to better tag PERSON-like legal signatories, ORG-like vendor names, and GPE/LOC-like addresses in semi-structured prose. Rule engine: A compact set of regexes implement detectors for emails (RFC 5322-guided), web URLs, IPv4/IPv6, phone numbers with optional country codes, and numeric identifiers with check-digit guards; each detection is replaced by a typed placeholder token preserving character offsets when feasible. Integration: A two-pass traversal applies regex first, then feeds masked text into spaCy to reduce spurious cues, with an aligner to map NER spans back to original offsets for consistent placeholder replacement and audit logs.datatracker.ietf+1

Evaluation Protocol

Entity-level metrics: Precision, recall, and F1 are computed at exact-span match for each category and macro-averaged across categories to compare regex-only, NER-only, and hybrid methods, conforming to de-identification shared-task practices. Redaction quality: BLEU and ROUGE between original and redacted-plus-placeholders documents assess structural and lexical preservation, where better scores imply minimal unintended text modifications outside sensitive spans, while still enforcing complete masking of targeted entities. Baselines: Regex-only (deterministic), spaCy-only (statistical), and a simple union baseline without overlap resolution are contrasted with the proposed precedence-aware hybrid pipeline to quantify additive benefits.pmc.ncbi.nlm.nih

Experimental Results

Overall performance: The hybrid approach yields higher macro F1 than regex-only and NER-only baselines, particularly improving recall on context-dependent entities (PERSON, ORG, GPE) while retaining high precision on pattern entities (EMAIL, URL, PHONE) via dedicated regex detectors. Content preservation: Documents redacted by the hybrid system achieve better BLEU/ROUGE against originals than the union baseline, reflecting fewer collateral edits and more consistent placeholdering, which maintains readability for reviewers and downstream analytics. Category analysis: Regex detectors dominate performance for EMAIL/URL/PHONE, while NER dominates for PERSON/ORG/GPE; the hybrid resolves ambiguous overlaps (e.g., organizational email handles) and reduces false positives triggered by capitalization heuristics.pmc.ncbi.nlm.nih

Comparative Analysis

Against regex-only: The hybrid significantly improves recall for entities lacking rigid formats, mitigating under-redaction risks that stem from purely pattern-based methods in narrative text. Against NER-only: The hybrid

10.48047/jocaaa.2019.27.07.09

reduces false negatives on semi-structured identifiers and minimizes span drift, since regexes offer deterministic boundaries, improving precision and edit locality. Relative to shared-task precedents emphasizing CRF or hybrid setups, the proposed implementation operationalizes a maintainable production stack with modern neural NER and standards-informed regexes to balance performance and engineering simplicity.

Discussion

Scalability: The design parallelizes over documents, with regex passes running in linear time and spaCy offering efficient batch processing; throughput scales across cores and containers for enterprise workloads such as large contract archives. Human-in-the-loop: An adjudication UI can surface low-confidence spans or category conflicts for reviewer verification, aligning with practical de-identification workflows demonstrated in prior shared tasks and institutional deployments. Error modes: Residual errors include unusual domain-specific identifiers and rare entity morphologies; targeted custom NER updates and incremental regex additions reduce these over time without destabilizing the pipeline.pmc.ncbi.nlm.nih

Conclusion

A hybrid redaction pipeline that unites regex-based detectors for pattern entities with spaCy NER for contextual entities delivers stronger privacy protection and document utility than single-technique baselines on adapted i2b2/UTHealth data. The approach supports typed placeholdering, scalable processing, and incremental domain adaptation via custom entity training, making it suitable for embedding alongside existing manual workflows in enterprise settings. Future work includes expanding domain-specific ontologies, refining overlap resolution policies, and extending evaluation to additional corpora and multilingual settings while preserving maintainability and auditability.spacy+1

References

- Stubbs, A., Filannino, M., Soysal, E., et al., "Overview of the 2014 i2b2/UTHealth Shared Task Track 1: De-identification of Longitudinal Clinical Narratives," Journal of Biomedical Informatics, 2015; includes corpus description, PHI categories, and system comparisons.pmc.ncbi.nlm.nih
- spaCy v2.0 documentation and model notes on improved NER with subword features and Bloom embeddings, indicating production-ready neural NER components circa 2017–2018.spacy+1
- Resnick, P., RFC 5322: Internet Message Format (October 2008), foundational for email address syntax referenced for practical regex design.datatracker.ietf
- i2b2 De-identification track resource page summarizing the 2014 task materials and availability for research evaluation under appropriate agreements.pubmed.ncbi.nlm.nih+1