Title: CONTEXT-AWARE SANITIZATION AND ONTOLOGICAL STRUCTURING OF UNSTRUCTURED DATA

Inventor: Dr. Michael E Hollins Jr

Address: 56 Beaver St., Apt. 205, New York, NY 10004 USA

Cross-Reference to Related Applications: None

BACKGROUND OF THE INVENTION

Field of the Invention

[0001] The present disclosure relates generally to the field of automated data processing systems. More particularly, the present disclosure relates to systems and methods that utilize machine learning models for the context-aware sanitization of sensitive information within unstructured text documents and the subsequent generation of structured, machine-readable metadata from the sanitized text using standardized ontologies.

Description of Related Art

[0002] The state of the art includes frameworks like 'RedactOR' by Singh et al. which teach de-identification using Large Language Models, and patents like U.S. Pat. No. 11,782,942 to Liang et al. which teach linking unstructured notes to structured data resources.

[0003] However, none of these references teach the specific, sequential process of the present invention, wherein the semantically preserved, sanitized text itself is used as the direct input for a programmatic data encoder to create a new layer of structured, ontology-mapped metadata. This unique step of structuring the privacy-preserved output solves the critical problem of creating a data asset that is simultaneously safe for analysis and structurally queryable.

BRIEF SUMMARY OF THE INVENTION

[0004] The present invention provides systems and methods for transforming unstructured text documents into privacy-preserving, structured data assets. The invention leverages a synergistic combination of a specialized machine learning model for context-aware redaction and a data encoder for mapping sanitized information to standardized ontologies. [0005] In one embodiment, a computer-implemented method comprises receiving an unstructured text document. A machine learning model, trained on a domain-specific corpus, processes the document to identify and redact sensitive text portions while preserving the grammatical and semantic context of surrounding text. Subsequently, a data encoder maps facts from the sanitized text to codes from a standardized ontology to generate structured metadata. The final output is a data object comprising both the human-readable sanitized text and the machine-readable structured metadata. [0006] In another embodiment, a system for performing

this method is provided, comprising a data ingestion module, an AI redaction engine, a data encoder, and a data storage interface.

BRIEF DESCRIPTION OF THE DRAWINGS

[0007] FIG. 1: System Architecture is a block diagram illustrating the overall system architecture.

[0008] FIG. 2: Method Flowchart is a flowchart illustrating the core method steps.

[0009] FIG. 3: AI Redaction Process Detail is a diagram detailing the process of sanitizing text while preserving semantic context.

[0010] FIG. 4: Output Data Object Structure is a diagram illustrating the exemplary data structure of the final output.

DETAILED DESCRIPTION OF THE INVENTION

[0011] The following detailed description is presented to enable any person skilled in the art to make and use the invention. For purposes of explanation, specific nomenclature and embodiments are set forth to provide a thorough understanding. However, it will be apparent to one skilled in the art that these specific details are not required to practice the invention and that the invention is not limited to the specific embodiments described herein. The descriptions of specific embodiments are provided for the purpose of illustration and representation, not limitation.

Terminology and Definitions

[0012] "Unstructured Text Document" refers to a document containing free-form human language, such as a clinical note or a legal contract, that does not have a predefined data model. [0013] "Sanitized Text Document" refers to a version of the unstructured text document where text portions corresponding to a predefined category of sensitive information have been redacted, while the grammatical structure and semantic context of the surrounding text are preserved. [0014] "Standardized Ontology" refers to a formal, standardized system of names, definitions, and categories that represents the concepts within a specific domain, such as healthcare, legal, or finance. [0015] "Data Object" refers to the final output of the system, a structured file comprising at least two distinct fields: one containing the human-readable sanitized text document and another containing the machine-readable structured metadata. [0016] "Context-Aware Redaction" refers to a process of identifying and modifying sensitive text portions within an unstructured text document in a manner that explicitly analyzes and preserves the grammatical structure, syntactical relationships, and overall meaning of the surrounding non-sensitive text, thereby maintaining linguistic coherence. [0017] "Grammatical and Semantic Context" refers to the relationships between words, phrases, and sentences within a text document that convey the intended meaning and allow for accurate interpretation. Preservation of this context ensures that even after redaction, the text remains logically coherent and suitable for subsequent automated natural language processing tasks. [0018]

"Predefined Category of Sensitive Information" refers to a classification of data types (e.g., names, dates, addresses, financial figures, medical diagnoses) explicitly designated by a system or user as requiring redaction to maintain privacy or confidentiality within an unstructured text document. [0019] "Identified Entities or Relationships" refers to specific concepts, named entities (e.g., persons, organizations, locations), or the connections between them (e.g., "patient suffers from disease," "party pays amount") extracted from the sanitized text document by the data encoder for mapping to a standardized ontology. [0020] "Computationally Queryable Data Asset" refers to a data object structured in such a way (e.g., containing both human-readable sanitized text and machine-readable structured metadata) that enables automated searching, filtering, analysis, and aggregation of information without requiring manual review or exposing sensitive underlying data. [0021] "Domain-Specific Corpus" refers to a collection of text documents, often large in size, that are specifically relevant to a particular subject area (e.g., medical records, legal contracts, financial reports) and used for training or fine-tuning machine learning models within that specialized field.

[0022] Referring now to FIG. 1, a block diagram illustrating the overall system architecture (100) is shown. The system comprises a Data Ingestion Module (110), an AI Redaction Engine (120), a Data Encoder (130), and a Data Storage Interface (140), which are communicatively coupled. For example, in a healthcare setting, the Data Ingestion Module (110) might receive thousands of daily clinical notes, discharge summaries, or pathology reports from various electronic health record (EHR) systems. In a legal context, it could ingest discovery documents, contracts, or court transcripts containing sensitive personally identifiable information (PII) or confidential business details.

[0023] It is a key technical advantage of the present architecture (100) that the Data Encoder (130) is positioned after the AI Redaction Engine (120) and is specifically configured to process the sanitized text output. A significant technical challenge in the art is that basic redaction can degrade or destroy the grammatical and semantic structures that a subsequent natural language processing module relies on to function accurately. The present invention solves this problem by ensuring the AI Redaction Engine (120) not only redacts sensitive information but is specifically configured to preserve the contextual integrity of the text for the express purpose of enabling high-fidelity, automated structuring by the Data Encoder (130). This synergistic, sequential processing is a non-obvious architectural choice that produces a technically superior data asset.

[0024] An additional key technical advantage of the present architecture is that the specific method of context-aware redaction performed by the AI Redaction Engine (120) is uniquely designed to support the subsequent structuring process. Prior art methods that rely on simple term generalization (e.g., replacing a specific entity with a generic placeholder like '[Company]' or '[Medication]') fundamentally destroy the granular semantic information that a high-fidelity information extraction process requires. Such methods would render a subsequent ontology mapping step either inoperable or highly inaccurate. The present invention solves this technical conflict by employing a sanitization technique that preserves the specific semantic concepts needed for the Data Encoder (130) to function with maximum precision, achieving a synergistic

result that would not be possible by simply combining unrelated sanitization and structuring tools.

[0025] The present invention decisively overcomes the limitations of existing de-identification and data structuring approaches by introducing a synergistic, sequential processing methodology. Prior art methods often degrade the linguistic integrity of text during redaction, thereby compromising the accuracy of subsequent automated structuring attempts. In contrast, the disclosed system introduces a context-aware AI Redaction Engine (120) precisely engineered to not only identify and redact sensitive information, but critically, to preserve the grammatical and semantic context of the surrounding text. This unique contextual preservation ensures that the sanitized text retains its interpretability for a subsequent Data Encoder (130). The Data Encoder (130) then programmatically maps entities and relationships from this linguistically coherent, privacy-preserved output to standardized ontologies, creating a new layer of highly accurate, machine-readable metadata. This non-obvious combination of advanced, context-preserving redaction directly feeding into structured metadata generation enables the creation of a novel data asset (400) that is simultaneously safe for analysis and structurally queryable, a critical advancement for fields handling sensitive unstructured information.

[0026] Referring now to FIG. 2, a flowchart illustrating the core method steps (200) is shown. The method begins at step 210 (Receive Unstructured Document). At step 220, a model identifies sensitive info. At step 230, a sanitized document is generated while preserving context. At step 240, a data encoder generates structured metadata. At step 250, a final data object (260) is stored. For instance, upon receiving a patient's clinical note (step 210), the AI Redaction Engine (120) identifies and redacts Protected Health Information (PHI) like patient names, addresses, and specific dates of service (step 220), creating a sanitized version (step 230). The Data Encoder (130) then extracts medical entities and relationships from this sanitized text, mapping them to standard healthcare ontology codes like ICD-10 or SNOMED CT (step 240). Finally, the combined sanitized text and structured codes are stored as a single, valuable data object (step 250) ready for secure research or analysis.

[0027] Referring now to FIG. 3, a diagram detailing the AI redaction process (300) is shown. An Input Text (310) (e.g., "On May 1st, 2024, ACME Corp. shall pay...") is received by the AI Redaction Engine (320). The engine redacts "ACME Corp." to produce an Output Text (330) (e.g., "On May 1st, 2024, [PARTY A] shall pay...") that preserves the original meaning. A more specific example for healthcare would be: Input Text (310): "Patient John Doe was admitted on 01/15/2023 with severe pneumonia and was discharged by Dr. Smith." The AI Redaction Engine (320) identifies "John Doe," "01/15/2023," and "Dr. Smith" as sensitive. The engine then generates Output Text (330): "Patient [PATIENT NAME] was admitted on [DATE] with severe pneumonia and was discharged by [PHYSICIAN NAME]." Notice how the surrounding grammatical and semantic context ("was admitted on," "with severe pneumonia," "was discharged by") remains entirely intact, allowing subsequent NLP processes to accurately understand the patient's condition and discharge details despite the redaction.

[0028] Referring now to FIG. 4, a diagram illustrating the output Data Object (400) is shown. The object comprises a Sanitized Text Field (410) and a Structured Metadata Field (420). The

Sanitized Text Field (410) contains the human-readable text, while the Structured Metadata Field (420) contains machine-readable JSON generated from an ontology. Continuing the healthcare example, the Sanitized Text Field (410) would hold the redacted clinical note. Concurrently, the Structured Metadata Field (420) would contain JSON data programmatically extracted from the sanitized note, perhaps including: {"diagnosis": "J18.9", "discharge_status": "home", "admitting_physician_specialty": "pulmonology"}. This structured metadata, linked to a standardized ontology like ICD-10 (for J18.9), allows researchers to efficiently query aggregated, de-identified data for trends (e.g., "How many patients with J18.9 were discharged to home last quarter?"), without directly accessing any PHI.

[0029] The foregoing description of the embodiments of the invention has been presented for the purposes of illustration and description. It is not intended to be exhaustive or to limit the invention to the precise form disclosed. Many modifications and variations are possible in light of the above teaching. It is intended that the scope of the invention be limited not by this detailed description, but rather by the claims appended hereto. The above specification, examples and data provide a complete description of the manufacture and use of the composition of the invention. Since many embodiments of the invention can be made without departing from the spirit and scope of the invention, the invention resides in the claims hereinafter appended.

[0030] It is to be understood that the systems and methods described herein may be implemented on one or more computing devices. Such a device will typically include at least one hardware processor and a memory storing instructions, which, when executed, cause the device to perform the steps of the methods described herein.

CLAIMS

What is claimed is:

1. A computer-implemented method for transforming an unstructured text document containing sensitive information, the method comprising: (a) receiving, at a hardware processor, the unstructured text document; (b) processing, by a machine learning model, the unstructured text document to identify one or more text portions corresponding to a predefined category of sensitive information; (c) generating a sanitized text document by redacting the identified one or more text portions, wherein the grammatical and semantic context of the surrounding text is preserved in the sanitized text document; (d) generating, by a data encoder, a set of structured metadata by programmatically mapping one or more identified entities or relationships from the sanitized text document to a corresponding set of codes from a standardized ontology; and (e) storing, in a memory, a data object comprising both the sanitized text document and the set of structured metadata.
2. The method of claim 1, wherein the machine learning model is a transformer-based machine learning model.

3. The method of claim 2, wherein the transformer-based machine learning model is fine-tuned on a domain-specific corpus of text documents.
4. The method of claim 1, wherein the predefined category of sensitive information is selected from the group consisting of Protected Health Information (PHI), Personally Identifiable Information (PII), and confidential financial data.
5. The method of claim 1, wherein the machine learning model is configured to identify the one or more text portions with high precision and recall, enabling an F1 score of greater than 0.95.
6. The method of claim 1, wherein the standardized ontology is selected from the group consisting of a healthcare ontology, a legal ontology, and a financial ontology.
7. The method of claim 1, wherein generating the sanitized text document comprises replacing the identified one or more text portions with a predefined placeholder.
8. The method of claim 1, wherein the structured metadata is generated in a JSON format.
9. The method of claim 1, wherein the data object is configured to facilitate programmatic querying of the structured metadata while maintaining the contextual integrity of the sanitized text document.
10. The method of claim 1, wherein the unstructured text document is selected from the group consisting of a clinical note, a legal contract, and a research paper.
11. The method of claim 1, wherein the data object is formatted for integration with an analytics platform or a secure data repository.
12. The method of claim 1, wherein the generating of the set of structured metadata further comprises: (a) performing semantic analysis on the sanitized text document to identify domain-specific concepts and relationships; (b) validating the mapped codes against the standardized ontology for consistency and completeness; and (c) generating confidence scores for each mapped code based on the semantic analysis, wherein only mappings exceeding a predetermined confidence threshold are included in the set of structured metadata.
13. A system for transforming an unstructured text document containing sensitive information, the system comprising: (a) a data ingestion module configured to receive the unstructured text document; (b) an AI redaction engine communicatively coupled to the data ingestion module, the AI redaction engine comprising a machine learning model configured to: (i) process the unstructured text document to identify one or more text portions corresponding to a predefined category of sensitive information; and (ii) generate a sanitized text document by redacting the identified one or more text portions while preserving the grammatical and semantic context of adjacent text; (c) a data encoder communicatively coupled to the AI redaction engine, the data encoder configured to generate a set of structured metadata by programmatically mapping one or more identified entities or relationships from the sanitized text document to a corresponding set of codes from a standardized ontology; and (d) a data storage interface configured to store a data object comprising the sanitized text document and the set of structured metadata.
14. The system of claim 13, wherein the machine learning model is a transformer-based machine learning model fine-tuned on a domain-specific text corpus.

15. The system of claim 13, wherein the predefined category of sensitive information is Protected Health Information (PHI).
16. The system of claim 13, wherein the machine learning model is configured to identify the one or more text portions with an F1 score of greater than 0.95.
17. The system of claim 13, further comprising an application programming interface (API) for accessing the stored data object.
18. The system of claim 13, further comprising a user interface for displaying the sanitized text document and the structured metadata.
19. The system of claim 13, wherein the data encoder further comprises: (a) a semantic analysis module configured to parse linguistic structures and extract domain-specific terminology from the sanitized text document; (b) a validation module configured to verify ontological mapping accuracy and detect potential mapping conflicts; and (c) a quality assurance module configured to generate metadata reliability metrics and filter low-confidence mappings, wherein the system maintains an audit trail of all mapping decisions for regulatory compliance.
20. A non-transitory computer-readable medium storing instructions that, when executed by one or more processors, cause the one or more processors to perform the method of any one of claims 1-12.

ABSTRACT

A system and method for transforming unstructured text documents into privacy-preserving, queryable data assets is disclosed. A context-aware machine learning model intelligently identifies and redacts sensitive information, uniquely preserving the critical grammatical and semantic context of the text. Subsequently, a data encoder programmatically maps entities and relationships from this sanitized text to standardized ontology codes, generating structured, machine-readable metadata. The resulting data object, comprising both the human-readable sanitized text and the structured metadata, creates a novel asset that is simultaneously safe for analysis and computationally queryable.