AI/ML Based Sensitive Data Discovery and Classification of Unstructured Data Sources

Shravani Ponde(✉), Akshay Kulkarni, and Rashmi Agarwal

RACE, REVA University, Bengaluru 560064, India
{shravanip.ba03,rashmi.agarwal}@reva.edu.in,
akshaykulkarni@race.reva.edu.in

Abstract. The amount of data produced every day is enormous. According to Forbes, 2.5 quintillion data is created daily (Marr, 2018). The volume of unstructured data is also multiplying daily, forcing organizations to spend significant time, effort, and money to manage and govern the data assets. This volume of unstructured data also leads to data privacy challenges in handling, auditing, and regulatory encounters thrown by governing bodies like Governments, Auditors, Data Protection/Legislative/Federal laws, regulatory acts like The General Data Protection Regulation (GDPR), The Basel Committee on Banking Supervision (BCBS), Health Insurance Portability and Accountability Act (HIPPA), The California Consumer Privacy Act (CCPA) etc.

Organizations must set up a robust data protection framework and governance to identify, classify, protect and monitor the sensitive data residing in the unstructured data sources. Data discovery and classification of the data assets is scanning the organization's data sources both structured and unstructured, that could potentially contain sensitive or regulated data.

Most organizations are using various data discovery and classification tools in scanning the structured and unstructured sources. The organizations cannot accomplish the overall privacy and protection needs due to the gaps observed in scanning and discovering sensitive data elements from unstructured sources. Hence, they are adapting to manual methodologies to fill these gaps.

The main objective of this study is to build a solution which systematically scans an unstructured data source and detects the sensitive data elements, auto classify as per the data classification categories, and visualizes the results on a dashboard. This solution uses Machine Learning (ML) and Natural Language Processing (NLP) techniques to detect the sensitive data elements contained in the unstructured data sources. It can be used as a first step before performing data encryption, tokenization, anonymization, and masking as part of the overall data protection journey.

Keywords: Data Discovery · Data Protection · Sensitive Data Classification · Data Privacy · Unstructured Data Discovery · Classification Model

S. Nandan Mohanty et al. (Eds.): ICISML 2022, LNICST 471, pp. 367–377, 2023.
https://doi.org/10.1007/978-3-031-35081-8_31

1 Introduction

The volume of the data owned by organizations is increasing daily, and data management is becoming a considerable challenge. CIO estimates that 80–90% of the data is in unstructured format (David, 2019). According to Forbes, 95% of businesses struggle to manage unstructured data (Kulkarni, 2019).

Meanwhile, data leakages, data breaches, and data security violations are also increasing drastically, which sometimes results in the organizations having to pay heavy penalties from the auditing and regulatory compliance aspects (Hill, 2022), which might also result in reputation loss.

1.1 Data Protection Laws and Regulations

Below are three pertinent Data Protection Laws:

The General Data Protection Regulation (GDPR)
European Union's (EU) GDPR is the law that imposes privacy regulations on any organization that accumulates or processes personal information related to individuals in the EU. Personal information includes but is not limited to names, email, location, ethnicity, gender, biometric data, religious beliefs, etc. All organizations are required to be GDPR compliant as of May 2018. The fines in case of GDPR violations are very high €20miilion or 4% of the global revenue (Wolford, 2020).

The California Consumer Privacy Act (CCPA)
The CCPA of 2018 gives Californian consumers control over how an organization collects their personal information. The personal information includes but is not limited to name, social security number, products purchased, internet browsing history, geolocation data, etc.

The CCPA provides consumers with three principal "rights." The first right is the "right to know" how the organization collects, uses, or shares personal information. The second right is the "right to opt-out" of selling personal data. The third right is the "right to delete" personal information collected about the consumer (Bonta, 2022).

The Health Insurance Portability and Accountability Act of 1996 (HIPAA)
HIPAA by the Department of Health and Human Services (HHS) gives consumers rights over their health information. Consumers have the right to get a copy of their health information, check who has it, and learn how it is used and shared. These regulations apply to health care providers, insurance companies, etc., (Office for Civil Rights (OCR), 2022).

Organizations are facing rapid growth of unstructured data, leading to the below challenges:

- Location of the unstructured data
- Classification per organization's policies
- Retention and disposal
- Monitoring of unstructured data

1.2 Data Discovery and Classification

Table 1 Gives a high-level overview of Data Discovery and Classification. It is very crucial to identify an organization's data assets scattered across the Enterprise. Organizations need to establish a robust data protection framework by defining security classification policies, Data Discovery methodologies, Data Privacy Standards and a practical Data Governance framework.

Table 1. Overview of Data Discovery and Data Classification.

Overview of Data Discovery and Data Classification	
Data Discovery	Data Classification
1. Identifying and Locating sensitive data in structured and unstructured sources via discovery rules 2. Identifying the data which is most at risk of exposure, such as PII, PHI	1. Categorizing the sensitive data - Internal, Public, Confidential, and Restricted 2. Classifying the sensitive data enables a faster search of the data assets across the enterprise

1.3 Data Protection Lifecycle

Organizations must identify, classify, protect and monitor sensitive data assets. To achieve this systematically, organizations need Data Protection Lifecycle (DPL) which helps organizations manage sensitive data. By accurately tracking sensitive data, organizations have a foundation to protect sensitive information and face future data privacy and protection challenges.

Figure 1 shows DPL, used to discover, classify, protect and protect sensitive data. By accurately tracking sensitive data, organizations have a foundation to protect sensitive information and face future data privacy and protection challenges.

1.4 Types of Sensitive Data

Sensitive data is confidential information that must be protected and inaccessible to outside parties. It can be of different types based on an organization's data classification policies (Steele, 2021):

- Personally Identifiable Information (PII)
- Sensitive Personal Information (SPI)
- Protected Health Information (PHI)
- Non-public Personal Information (NPI)

Table 2 Provides examples of different types of sensitive data.

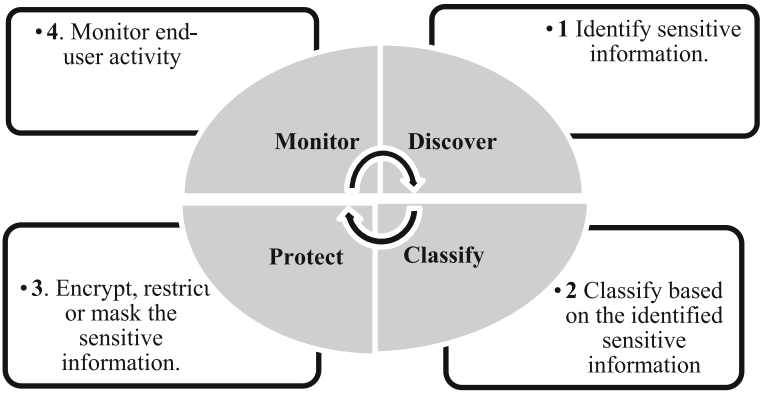

Fig. 1. Data Protection Lifecycle

Table 2. Types of Sensitive Data.

Types of Sensitive Data			
PII	SPI	PHI	NPI
1. Name 2. e-mail address 3. Phone Number 4. Date of Birth 5. Address	1. SSN 2. Driver License 3. Passport Number 4. Religious beliefs 5. Political Opinion 6. Genetic data 7. Biometric data	1. Medical Information 2. Physical Health Information 3. Mental Health Information	1. Bank Account Number 2. Credit Card Number

2 Literature Review

Data is considered capital in today's digital economy and holds tremendous value. "Data is regarded as the new oil," said Clive Humby. Organizations are increasingly relying on a robust data management strategy to use data and create value. One of the critical aspects of data management is to manage sensitive data across the enterprise. (Goswami, 2020) states that 69% of consumers are concerned about how personal data is collected in mobile apps. (Gartner Top Strategic Technology Trends for 2022, 2022) lists' Privacy-enhancing computation techniques' as one of the top technology trends for 2022. As per Gartner securing personal data is critical due to evolving privacy and data protection laws and growing consumer concerns. (Yaqoob, Salah, Jayaraman, & Al-Hammadi, 2022) outline data privacy as one of the critical challenges to healthcare data management.

As per Oracle (What Is Data Management?, 2022), today's organizations' data management systems include databases, data lakes, data warehouses, the cloud, etc. Big data management systems have emerged as more and more data is collected every day from sources as disparate as video cameras, social media, audio recordings, and Internet of Things (IoT) devices. Compliance regulations are complex and multijurisdictional, and

they change constantly. Organizations need to be able to review their data quickly; in particular, personally identifiable information (PII) must be detected, tracked, and monitored for compliance with increasingly strict global privacy regulations. (Mehmood, Natgunanathan, Xiang, Hua, & Guo, 2016) illustrate the infrastructure of big data and the privacy-preserving mechanisms in each stage of the big data life cycle.

This solution focuses on the capabilities of data governance and data management framework. The framework establishes, enables, and sustains a mature data privacy management solution, which is the core discipline in the data management and governance arena. In (Cha & Yeh, 2018) proposed a data-driven risk assessment approach to personal data protection, which can prevent organizations from overlooking risks to sensitive data. In (Truong, Sun, Lee, & Guo, 2019) design a concept for GDPR compliant Block Chain based personal data management solution. (Xu, Jiang, Wang, Yuan, & Ren, 2014), discusses the approach to privacy protection and proposes a user role-based methodology to privacy issues. (Zhang, et al., 2017) Propose a scalable MRMondrian approach for multidimensional anonymization over big data based on the MapReduce paradigm.

3 Problem Statement

Organizations leverage unstructured data across the enterprise, which results in an ever-increasing volume of data that requires protection. A complete data lifecycle is necessary to manage data from its creation to its destruction, ensuring that appropriate protections are applied along the way.

Organizations are scrutinized as to how they manage, control, and monitor stakeholders' data and their preferences. As data breaches increase and sensitive information is compromised, more privacy regulations are developed, from state/ national requirements to potential comprehensive federal privacy laws.

Global privacy legislations require the clients to document and take responsibility for personal data and processing activities. The data discovery and classification program can help them to comply with this requirement.

Below are some of the benefits of sensitive data discovery and classification of unstructured sources:

- Visibility to the sensitive data
- Reduced sensitive data footprint that is not needed
- Enhanced governance and protection of data when stored and transferred internally and externally
- Integrations with data loss preventions, information rights management, defender for end points
- Maintain compliance, apply risk-based protections

Organizations can protect sensitive data if they know where it resides. The data discovery and classification help clients identify where sensitive data is stored and enable the application of risk-based protections.

It is crucial to identify, classify and protect the sensitive data to drive the below initiatives for the organizations as applicable:

- Regulatory Compliance – GDPR, CCPA, etc.

- Auditing purposes
- Data Privacy and protection needs for customers, employees, suppliers, etc.,
- Data governance
- Enterprise metadata management
- Data Remediation
- Data Disposals
- Data Subject Rights

4 Proposed Solution

Solve one of the primary data privacy challenges discussed – help organizations manage the sensitive data on unstructured files stored across – Confluence, SharePoint, shared network drives, etc.

a) Detect sensitive elements
b) Document Risk Categorization
c) Document Classification

4.1 Detect Sensitive Elements

A sensitive PII data element has three parts – format, pattern, and keywords.

Format: Format of the sensitive data element.

Pattern: The sensitive data elements are pattern-based classifiers that can be identified using regular expressions. A pattern defines what the sensitive data element looks like.

Keywords: Keywords are used to identify the sensitive data element. They represent the occurrences of sensitive data elements in the unstructured data source.

Social Security Number

USA Social Security Number (SSN) consists of 9 digits. The first set of three digits is called the Area Number. The second set of two digits is called the Group Number. The final set of four digits is the Serial Number. Table 3 shows the format and sample for identifying an SSN. Table 3 shows the pattern and sample for identifying an SSN.

Table 3. Sensitive Data Element – SSN.

Sensitive Data Element		
SSN	Pattern	Sample
	ddd-dd-dddd	986–43-2453
	ddd dd dddd	231 24 3168

Format: Nine digits

Pattern: Search the pattern with formatting that has dashes or spaces (ddd-dd-dddd OR ddd dd dddd)

Table 4. Sensitive Data Element – e-mail address.

Sensitive Data Element		
e-mail address	Pattern	Sample
	< Letters > @ < letters >. < letters >	Mark.Campbell@gmail.com Lisa.Thomas@hotmail.com

Keywords: ssn, social security number, ssn n, social security #, social security no, Soc sc, s#

E-mail Address

Format: Search the pattern with letters followed by '@' and '.'

Keywords: email, e-mail, email address, e-mail address, email id etc.

The format and pattern for other sensitive elements like Name, Phone number, Date of birth etc., was defined similarly. Regular Expressions (RegEx) are used to identify the sensitive data elements.

4.2 Rule Based Document Risk Categorization

The sensitive data elements identified in a document using RegEx (from Sect. 4.1), is used for the document risk categorization. The Table 5 Illustrates the rule-based document categorization approach followed. For example, if a document contains a name along with SSN, PAN, or DOB is categorized as a high-risk document. If a document includes either SSN, Phone number, or e-mail address is categorized as a medium-risk document. If a document contains only a name or DOB is categorized as a low-risk document.

Table 5. Rule based Risk Categorization Matrix

Sensitive Data Element	Document Risk Categorization		
	High	Medium	Low
Name			Yes
SSN		Yes	
DOB			Yes
Phone Number		Yes	
email		Yes	
Name + SSN	Yes		
Name + Phone Number		Yes	
Name + email		Yes	
Name + DOB	Yes		

4.3 Document Classification

Typically, there are four classifications of data. A document can be classified as Public, Internal, Confidential, or Restricted.

Public
This type of data is freely accessible to the public.

Internal
This type of data is strictly for internal company personnel.

Confidential
This type of data is sensitive, and only selective access is granted.

Restricted
This type of data has proprietary information and needs the authorization to access it. Inappropriate handling can lead to criminal or civil charges.

Table 6 shows the type of information contained in each document category. For example, an organization's public document can contain financial statements, press releases, etc. In contrast, a restricted document might contain sensitive information like SSN or Bank Account Numbers.

Table 6. Document Category.

Document Category			
Public	Internal	Confidential (Non-Sensitive PII)	Restricted (Sensitive PII)
1. Financial Statements 2. Press Release	1. Training Materials 2. Instructions	1. Name 2. Phone Number 3. e-mail address	1. SSN 2. Date of Birth 3. Bank Account Number

Synthetic Data Generation
Since sensitive information (PII) is unavailable on open sources, synthetic data which mimics PII (Restricted and Confidential) was generated while preserving the format and data type.

Text Pre Processing
The unstructured word documents are cleaned and pre-processed to make it ready for modelling. First the text is standardized by converting to lowercase. All the special characters, numbers and stop words are removed. Lemmatization is used to return the base or dictionary form of words (lemma). In this step we transform the words into their normalized form. Figure 2 shows the text cleaning pipeline used for pre-processing the documents.

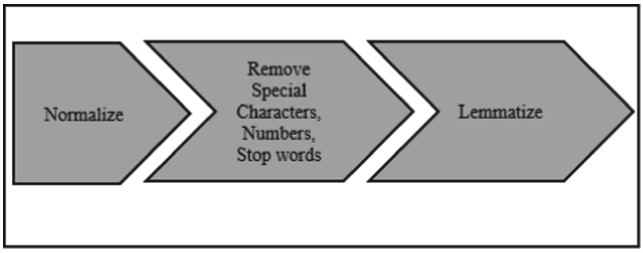

Fig. 2. Text pre-processing pipeline

Feature Engineering

To analyze a preprocessed data, it needs to be converted into features. Under Feature Engineering, features are created from the cleaned text so that the machine learning model can be trained as shown in Fig. 3.

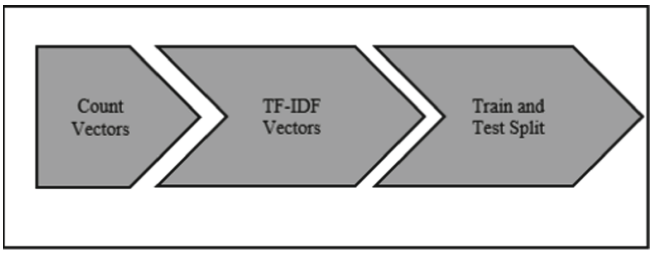

Fig. 3. Feature engineering pipeline

4.4 Data Modelling Results

Multiclass classification (multinomial classification) refers to supervised machine learning with classification of more than two classes. Since, the documents belong to more than one category, multi-class classification algorithm was used to auto-classify the document.

Table 7 shows the data modelling results for the various classifiers. The best model based on test accuracies is Multinomial Naïve Bayes. After the model training process, the trained model is saved and used to classify the document.

Table 7. Classification Data Modelling Results.

Data Modelling Results		
Classifier	Count Vectors Accuracy	TF-IDF Accuracy
Multinomial Naïve Bayes	90%	62%
Random Forest	80%	68%
K Neighbours	68%	50%
Decision Tree	83%	56%

5 Conclusion and Future Scope

This solution be customized according to privacy and classification needs and deployed on on-premise and cloud infrastructure platforms via APIs.

After the Sensitive data discovery and classification of the sensitive data elements, organizations can commence the below data privacy and data protection needs, and the organizations would establish strong data protection and governance framework to handle regulatory and auditing challenges well in advance.

Enabling access and security controls

- Role based access control
- Password protection control
- API level Encryptions

Data Protection Capabilities

- Data Masking, Encryption, Tokenization
- Anonymization, Pseudonymization
- Data Loss Prevention (DLP)
- Data Remediation

Data Subject Rights.

References

Bonta, R.: California Consumer Privacy Act (CCPA). Retrieved from State of California Department of Justice: https://oag.ca.gov/privacy/ccpa (2022)

Cha, S.-C., Yeh, K.-H.: A Data-Driven Security Risk Assessment Scheme for Personal Data Protection. IEEE, pp. 50510 – 50517 (2018)

David, D.: AI Unleashes the Power of Unstructured Data. Retrieved from CIO (2019, July 9). https://www.cio.com/article/3406806/ai-unleashes-the-power-of-unstructured-data.html

Gartner Top Strategic Technology Trends for 2022. (2022). Retrieved from Gartner: https://www.gartner.com/en/information-technology/insights/top-technology-trends

Goswami, S.: The Rising Concern Around Consumer Data And Privacy. Retrieved from Forbes (2020, December 14). https://www.forbes.com/sites/forbestechcouncil/2020/12/14/the-rising-concern-around-consumer-data-and-privacy/?sh=30741b43487e

Hill, M.: The 12 biggest data breach fines, penalties, and settlements so far. Retrieved from CSO (2022, August 16). https://www.csoonline.com/article/3410278/the-biggest-data-breach-fines-penalties-and-settlements-so-far.html

Kulkarni, R.: Big Data Goes Big. Retrieved from Forbes (2019, 02 07). https://www.forbes.com/sites/rkulkarni/2019/02/07/big-data-goes-big/?sh=278b2aa820d7

Marr, B.: How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read. Retrieved from Forbes (2018, May 21). https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/?sh=4e4f805860ba

Mehmood, A., Natgunanathan, I., Xiang, Y., Hua, G., Guo, S.: Protection of Big Data Privacy. IEEE, pp. 1821–1834 (2016)

Office for Civil Rights (OCR). (2022, January 19). Your Rights Under HIPAA. Retrieved from HHS.gov: https://www.hhs.gov/hipaa/for-individuals/guidance-materials-for-consumers/index.html

Steele, K.: A Guide to Types of Sensitive Information. Retrieved from BigID (2021, November 3). https://bigid.com/blog/sensitive-information-guide/

Truong, N.B., Sun, K., Lee, G.M., Guo, Y.: GDPR-Compliant Personal Data Management: A Blockchain-Based Solution. IEEE, pp. 1746–1761 (2019)

What Is Data Management? (2022). Retrieved from OCI: https://www.oracle.com/database/what-is-data-management/

Wolford, B.: What is GDPR, the EU's new data protection law? Retrieved from GDPR.EU (2020) https://gdpr.eu/what-is-gdpr/

Xu, L., Jiang, C., Wang, J., Yuan, J., Ren, Y.: Information Security in Big Data: Privacy and Data Mining. IEEE, pp. 1149–1176 (2014)

Yaqoob, I., Salah, K., Jayaraman, R., & Al-Hammadi, Y.: Blockchain for healthcare data management: opportunities, challenges, and future recommendations. Springer Link, pp. 11475–11490 (2022)

Zhang, X., et al.: MRMondrian: Scalable Multidimensional Anonymisation for Big Data Privacy Preservation. IEEE, pp. 125–139 (2017)