REVA UNIVERSITY
Bengaluru, India

A Project Report on

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

Submitted in Partial Fulfilment for Award of Degree of
Master of Business Administration
In Business Analytics

Submitted By
Shravani Ponde
R18DMO18

Under the Guidance of
Akshay Kulkarni
Manager Data Science & AI, Publicis Sapient

REVA Academy for Corporate Excellence - RACE
REVA University
Rukmini Knowledge Park, Kattigenahalli, Yelahanka, Bengaluru - 560 064
race.reva.edu.in

August, 2022

REVA UNIVERSITY
Bengaluru, India

Candidate's Declaration

I, Shravani Ponde hereby declare that I have completed the project work towards the second year of Master of Business Administration in Business Analytics at, REVA University on the topic entitled **AI/ML Based Sensitive Data Discovery and Classification of Unstructured Data Sources** under the supervision of Mr. Akshay Kulkarni. This report embodies the original work done by me in partial fulfilment of the requirements for the award of degree for the academic year 2022.

Place: Bengaluru
Date: 27-08-2022

Name of the Student: Shravani Ponde
Signature of Student: Shravani Ponde

REVA UNIVERSITY
Bengaluru, India

Certificate

This is to Certify that the Project work entitled **AI/ML Based Sensitive Data Discovery and Classification of Unstructured Data Sources** carried out by Shravani Ponde with R18DMO18, is a bonafide student of REVA University, is submitting the second year project report in fulfilment for the award of MBA in Business Analytics during the academic year 2022. The Project report has been tested for plagiarism, and has passed the plagiarism test with the similarity score less than 15%. The project report has been approved as it satisfies the academic requirements in respect of project work prescribed for the said Degree.

Akshay Kulkarni

Signature of the Guide
Akshay Kulkarni
Guide

Signature of the Director
Dr. Shinu Abhi
Director

External Viva
Names of the Examiners

1. Dr. Sai Hareesh, Research Expert, SAP Labs India
2. Pradeepta Mishra, Director – AI, L&T InfoTech

Place: Bengaluru
Date: 27-08-2022

Acknowledgement

I would like to thank my mentors, trainers, classmates, program office members at REVA University for their continuous support and guidance.

I would also like thank my family and friends who have directly and indirectly supported me in this project.

I would like to thank Hon'ble Chancellor, Dr. P Shayma Raju, Hon'ble Vice Chancellor, Dr. M. Dhanamjaya, and Registrar, Dr. N. Ramesh for the opportunity and providing the essential resources.

Place: Bengaluru
Date: 27-08-2022

REVA UNIVERSITY
Bengaluru, India

Similarity Index Report

This is to certify that this project report titled **AI/ML Based Sensitive Data Discovery and Classification of Unstructured Data Sources** was scanned for similarity detection. Process and outcome is given below.

Software Used: Turnitn
Date of Report Generation: 26-08-2022
Similarity Index in %: 9%
Total word count: 5938
Name of the Guide: Akshay Kulkarni

Place: Bengaluru
Date: 27-08-2022

Shravani Ponde
Signature of Student: Shravani Ponde

Verified by:

Signature
Dr. Shinu Abhi,
Director, Corporate Training

List of Abbreviations

Sl. No	Abbreviation	Long Form
1	AI	Artificial Intelligence
2	ML	Machine Learning
3	NLP	Natural Language Processing
4	PII	Personal Identifiable Information
5	SPI	Sensitive Personal Information
6	PHI	Protected Health Information
7	NPI	Non-public Personal Information
8	GDPR	General Data Protection Regulation
9	BCBS	The Basel Committee on Banking Supervision
10	CCPA	The California Consumer Privacy Act of 2018
11	CCAR	The Comprehensive Capital Analysis and Review
12	HIPAA	Health Insurance Portability and Accountability Act
13	MDM	Master Data Management
14	DPM	Data Privacy Management

List of Figures

List of Tables

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 (Governments, Auditors, Data Protection/Legislative/Federal laws, etc.) and regulatory acts (GDPR, BCBS, Hippa, CCPA, CCAR, 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 (structured and unstructured) that could potentially contain sensitive or regulated data.

Most organizations are using various data discovery and classification tools and technologies (Informatica, IBM, Alation, BigID, MIP, Google, etc.) 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 adopting manual methodologies to fill these gaps.

The main objective of this project is to build a model 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 model uses advanced AI/ML and Natural Language Processing (NLP) techniques to detect the sensitive data elements across the different types of unstructured sources.

This model can be deployed and customized to detect, and auto classifies any data, both on-premise and in cloud environments. This model 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, Data tagging, Data labelling, Unstructured Data Discovery, Classification Model

Table of Contents

Chapter 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 & Regulations

Below are three pertinent Data Protection Laws:

1.1.1 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)**.

1.1.2 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)**.

1.1.3 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)**.

1.2 Key Drivers for Regulatory Compliance:

The below critical drivers drive the organization to be regulatory compliant concerning various regulatory compliance acts and programs held by government authorities on managing the customers' data across all types of enterprises and adherence to laws, regulations, guidelines, and specifications relevant to their business processes.

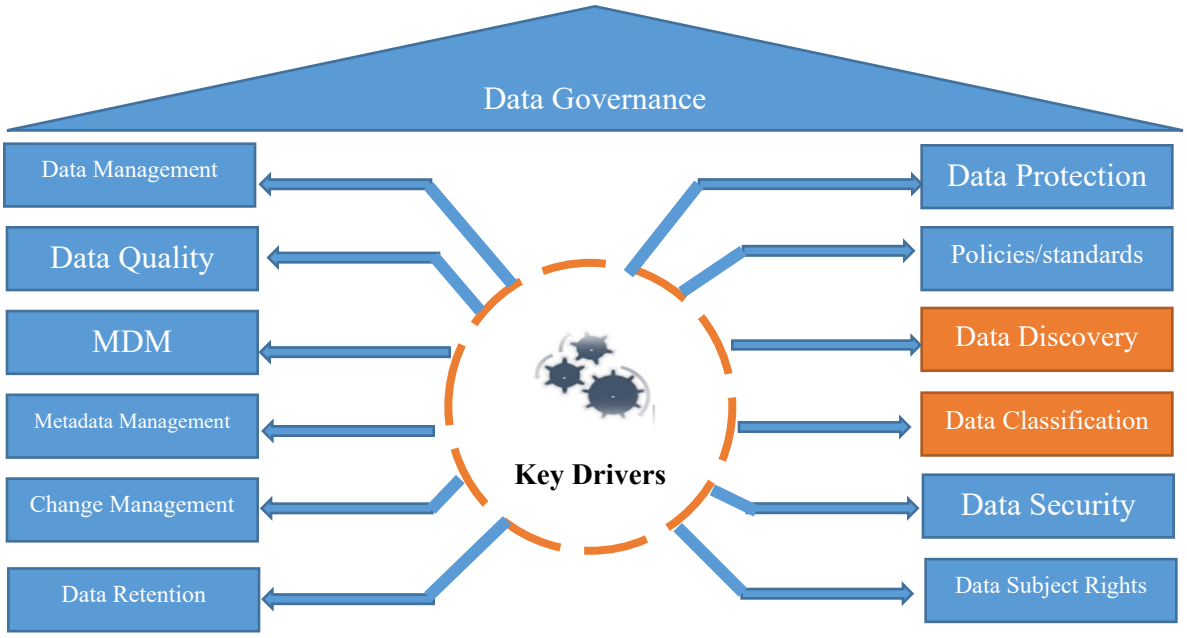


Figure No. 1.1: Key Drivers for Regulatory Compliance

This project focuses on **data discovery** and **classification** of unstructured data sources, which helps organizations discover sensitive data assets and auto-classify the unstructured documents based on the classification policies and data standards.

1.2.1 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, etc.,	1. Categorizing the discovered sensitive data across various sensitivity levels (Internal, Public, Confidential, and Restricted). 2. Auto-classifying sensitive data enables a faster search of the data assets across the enterprise.

TABLE NO. 1.1 OVERVIEW OF DATA DISCOVERY & DATA CLASSIFICATION

Table No. 1.1 gives a high-level overview of Data Discovery & Data 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, a robust execution model, and a practical Data Governance framework.

1.3 Data Protection Lifecycle (DPL)

Organizations must understand the data assets and categorize, protect and monitor sensitive data assets. To achieve this systematically, organizations need a sophisticated data discovery and classification of the data assets. Data protection lifecycle, or DPL, helps organizations to manage sensitive data.

Figure No. 1.2 shows the 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.

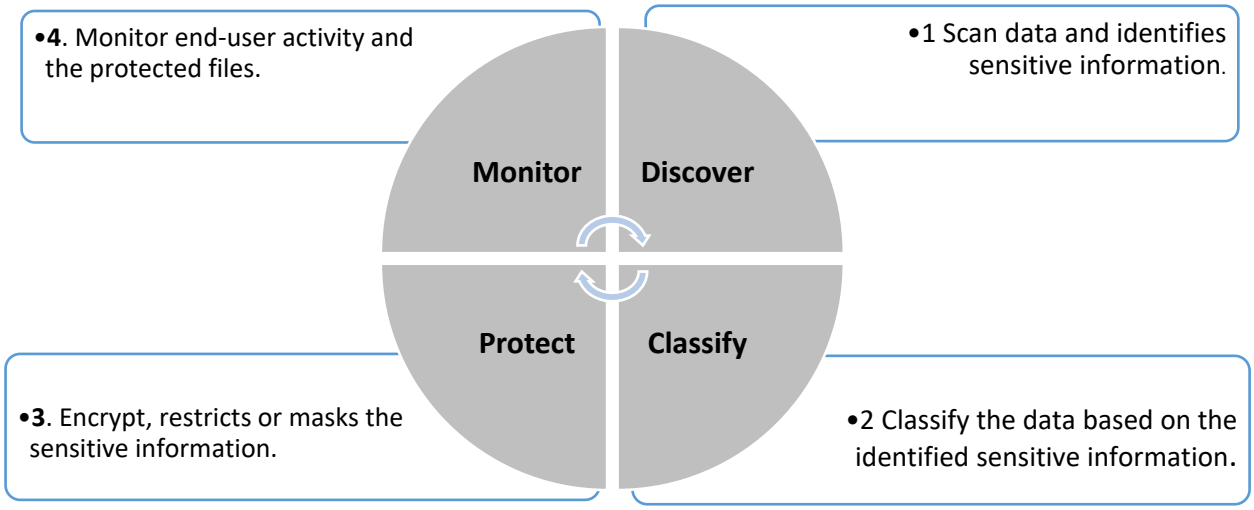


Figure No. 1.2: Data Protection Life Cycle

1.4 Sensitive Data

Sensitive data is confidential information that must be protected and inaccessible to outside parties. Sensitive data can be in physical or electronic form; it is private information. It is crucial to locate sensitive data because:

1. To protect sensitive data from untrusted users and security violators.
2. To avoid data loss, unexpected data breaches, and misuse of data both internally and externally in an organization.
3. To provide hassle-free auditing reports on time and comply with all the security aspects.

Depending on the industry, an organization will have to comply with more than one regulation. For example, a communications industry has to be compliant with both GDPR and CCPA.

1.4.1 Types of Sensitive Data

Sensitive data can be of different types based on an organization's data classification policies (Steele, 2021):

1. Personally Identifiable Information (PII)
2. Sensitive Personal Information (SPI)
3. Protected Health Information (PHI)
4. Non-public Personal Information (NPI)

Figure No. 1.3 provides examples of each type of Sensitive Data.

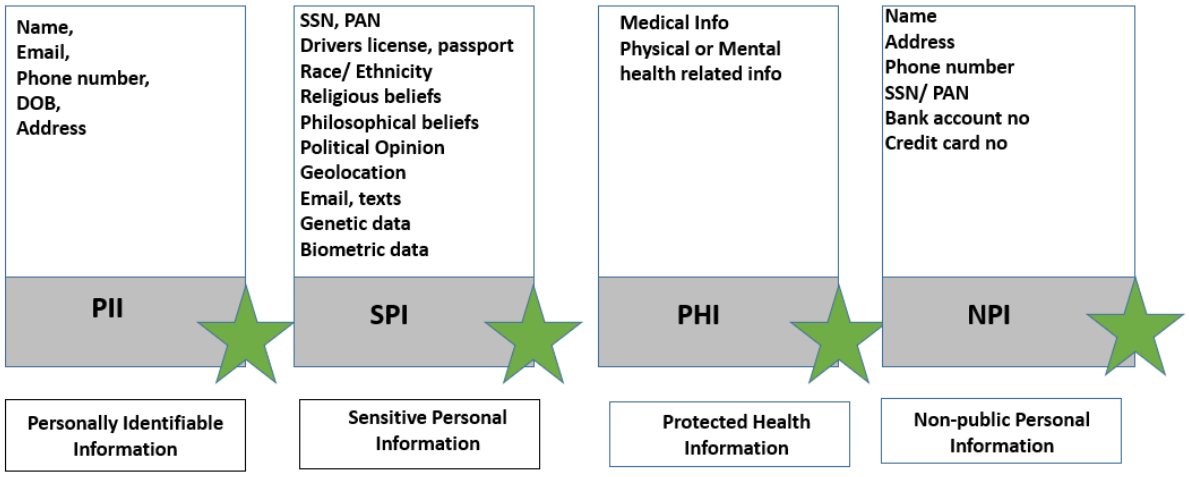


Figure No. 1.3: Types of Sensitive Data

1.5 Structured & Unstructured Data sources

Data is categorized as structured, unstructured, and semi-structured:

Structured data

Structured data is information that is organized in a standard format using a predefined data model stored in Relational Databases (RDBMS), Data warehouses, etc.

Unstructured data

Unstructured data is information that is not arranged according to a predefined data model stored in NoSQL databases, data lakes, shared drives, confluence, etc.

Examples of unstructured data:

- Word documents
- PowerPoint presentations
- Web pages
- Images (jpeg, gif, png)
- Videos
- Emails
- Surveys

1.6 Benefits of sensitive data discovery and classification

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.

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

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

Chapter 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 project focuses on the capabilities of data governance and data management framework. The project 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.

(Gai, Qiu, & Zhao, 2017) propose an approach to selectively encrypt data and use privacy classification methods under timing constraints. Their system is to encrypt data and use privacy classification methods.
This model built as a part of this project can handle sensitive data discovery and classification of unstructured documents. The model can scan unstructured documents like word, pdf, etc., detect the sensitive PII data elements, auto classify them based on the sensitivity levels, and calculate the risk scores.

Chapter 3: Problem Statement

3.1 Problem Statement

Organizations are facing rapid growth of unstructured data, leading to the below gaps, concerns, and challenges. Below are the Observed gaps that most organizations are facing today:

1. As per the annual privacy governance reports, most firms struggle to “locate unstructured personal data.”
2. The majority of the firms have not performed comprehensive work on unstructured data analysis.
3. Lot of Privacy initiatives are on hold due to the unavailability of automated processes to identify and classify sensitive data elements from unstructured sources.

3.2 Common concerns and challenges:

Firms are expressing common concerns around managing unstructured data, data minimization, and deletion, such as:

1. Automatic location of the unstructured data.
2. Classification of unstructured data per organization's policies and standards.
3. Retention, Disposal/deletion of unstructured data per the policies.
4. Monitoring of unstructured data per the policies.

Chapter 4: Objectives of the Study

Solve one of the primary data privacy challenges discussed in Chapter 3 – help organizations manage the sensitive data on unstructured files stored across – Confluence, SharePoint, shared network drives, etc.

The main objectives of this project:

1. Apply AI/ML and Natural Language Processing (NLP) techniques to systematically scan the unstructured document and detect the sensitive data elements.
2. Auto classify the unstructured documents per the organization's data classification policies – Public, Internal, Confidential, Restricted. Rule-based risk categorization of the document into one of the categories – Low, Medium, High.
3. Create a prototype of a scalable solution that can handle huge volumes and other unstructured data types across on-premise and cloud platforms.

Chapter 5: Project Methodology

There are four main parts to this project, as shown in Figure No 5.1:

5.1 Detect the sensitive PII data elements, calculate the document risk score

5.2 Auto classify

5.3 Document Risk Categorization

5.4 Visualize the results

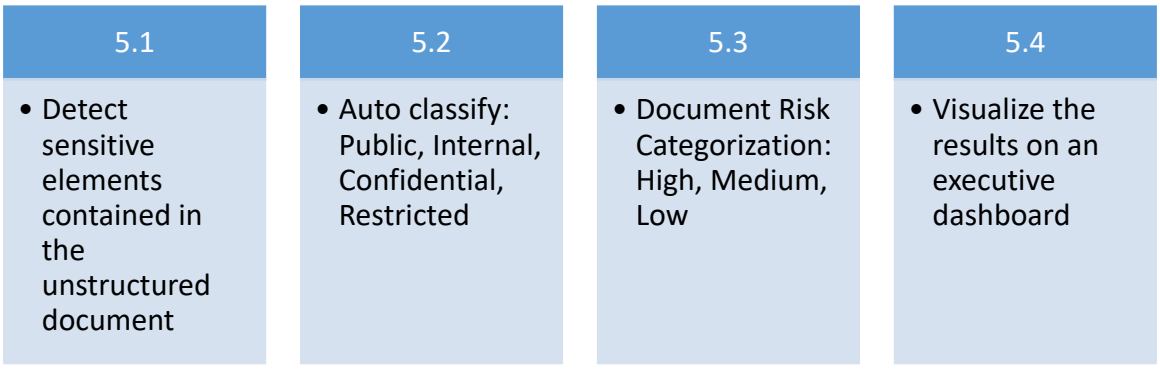


Figure No. 5.1: Project Methodology

5.1 – Detect sensitive PII data elements contained in the document

Sensitive PII data elements contained in the word document are identified using regular expressions. 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.

a) Person Name

Table No. 5.1.1 shows the format and sample for identifying a name.

Sensitive Data Element	Format	Sample
Name	First Name, Last Name	Mark Campbell Lisa Thomas

TABLE NO. 5.1.1: PERSON'S NAME – FORMAT/ SAMPLE

Format

First Name, Last Name

Keywords

Name, full name, full_name, first name, last name

b) Email Address

Table No. 5.1.2 shows the format and sample for identifying an email address.

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

TABLE NO. 5.1.2: E-MAIL ADDRESS – PATTERN/ SAMPLE

Format

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

Keywords

email, e-mail, email address, e-mail address, email id, e-mail id

c) Date of Birth

Table No. 5.1.3 shows the format and sample for identifying the date of birth.

Sensitive Data Element	Country	Pattern	Sample
Date of Birth	United States	MM-DD-YYYY MM DD YYYY	06-25-1978 02 22 1975
	India	DD-MM-YYYY DD MM YYYY	10-03-1985 30 08 1986

TABLE NO. 5.1.3: DATE OF BIRTH – PATTERN/ SAMPLE

Format

Eight digits

Pattern - US

Search the pattern with eight digits and two dashes or spaces in the format MM-DD-YYYY or MM DD YYYY.

Pattern - India

Search the pattern with eight digits and two dashes or spaces in the format DD-MM-YYYY or DD MM YYYY.

Keywords

Dob, d.o.b, date of birth, birth date, birth_date

d) Phone number

Table No. 5.1.4 shows the format and sample for identifying a phone number.

Sensitive Data Element	Country	Pattern	Sample
Phone Number	India - IN	91<Numbers>	919443452187
	United States - US	1<Numbers>	14699562467

TABLE NO. 5.1.4: PHONE NUMBER – PATTERN/ SAMPLE

Pattern

Ten numbers with a prefix of 91.

Ten numbers with a prefix of 1.

Keywords

Number, phone, phone number, phone no, phone#, mobile, mobile number, mobile no, mobile#, contact, contact number, contact no, contact#

e) India Permanent Account Number (PAN)

Table No. 5.1.5 shows the format and sample for identifying an Indian PAN.

Sensitive Data Element	Pattern	Sample
Permanent Account Number	Five Letters Four Numbers One letter	xscpp2818e

Table No. 5.1.5: PAN – Pattern/ Sample

Format

10 letters or digits

Pattern

First five letters

Next four numbers

Last a letter

Keywords - PAN

Permanent Account Number, PAN

f) U.S. social security number (SSN)

Table No. 5.1.6 shows the format and sample for identifying a US SSN.

Sensitive Data Element	Pattern	Sample
Social Security Number	ddd-dd-dddd ddd dd dddd	986-43-2453 231 24 3168

TABLE NO. 5.1.6: SSN – PATTERN/ SAMPLE

Format

Nine digits

Pattern

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

Keyword - SSN

ssn, social security number, ssn number, social security #, social security no, soc sec, ssn#

5.2 - Auto classify the word document

The word document is classified as Public, Internal, Confidential, or Restricted using a multi-class classification model. Figure No. 5.2 shows the document classification pipeline

Figure No. 5.2: Document Classification Pipeline

Scope – Word documents

Identify the below sensitive data elements from the unstructured word documents.

1. Name
2. Phone number
3. E-mail address
4. Date of birth
5. Social security number/ Permanent account number

Generate Sensitive Data Samples

Synthetic data is used to demonstrate the output, which gives a realistic sense of discovering and classifying the sensitive data.

Develop the multi class classification model

Classification Algorithm is used to predict the class of the document. Since, the documents have more than one category, multi-class classification algorithm is used for document classification. Some of the popular multi-class classification algorithms like K-Nearest Neighbors, Decision Trees, Multinomial Naïve Bayes were used.

Execution

The data is cleaned following a text cleaning pipeline and split into two datasets – Train and Test. The training set is used by the model to learn about the data. The different machine learning algorithms are then trained on the Train dataset.

Test and validate

After training the model, the test dataset is then used to test the accuracy of the model. The different models are then evaluated using the results from the test dataset.

5.3 Document Risk Categorization

The Table No. 5.7 illustrates the Risk Categorization criteria:

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, PAN, 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.

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

TABLE NO. 5.7 : DOCUMENT RISK CATEGORIZATION

Chapter 6: Business Understanding

Organizations must protect their stakeholder's privacy and keep their data secure by adhering to privacy compliance. As the number of data transactions grows, organizations must manage more data. Plus, the data privacy compliance rules are continuously changing in which unstructured data sources are increasing, and the technologies aren't supportive of scanning them and protecting sensitive information systematically.

Depending on the sensitivity of the data, there are different classification.
Typically, there are four classifications of data:

1. **Public**
   This type of data is freely accessible to the public.
2. **Internal**
   This type of data is strictly for internal company personnel.
3. **Confidential**
   This type of data is sensitive, and only selective access is granted.
4. **Restricted**
   This type of data has proprietary information and needs the authorization to access it. Inappropriate handling can lead to criminal or civil charges.

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.

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:

1. Regulatory Compliance – GDPR, CCPA, etc.
2. Auditing purposes
3. Data Privacy and protection needs for customers, employees, suppliers, etc.,
4. Data governance
5. Data quality and MDM
6. Enterprise metadata management
7. Data Remediation
8. Data Disposals
9. Data Subject Rights

This solution can help organizations fill the gap in discovering and classifying sensitive data across unstructured sources. Which can help the organizations build brand trust, better manage risk, and avoid non-compliance fines, reputation loss, and future opportunity loss.
Since this project is designed on an open-source platform, it is easy to integrate on any platform at a cheaper cost. It can be customized and scaled as per the auditing and regulatory needs of any organization.

Chapter 7: Data Understanding

Identifying and classifying the data assets scattered across the enterprise is an integral part of the data protection lifecycle. Synthetic data generation is based on sensitive data elements classified across these four levels. Table No. 7.1 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.

Document Category			
Public	Internal	Confidential	Restricted
		Non-Sensitive PII:	Sensitive PII:
Financial Statements	Training Materials	Name	SSN/ PAN
Press Release	Instructions	Phone Number	Date of Birth
		e-mail address	Bank Account Number

TABLE NO. 7.1: DOCUMENT CATEGORY

The classification model input is unstructured data in various documents across the different document categories as per the Figure No. 7.1.

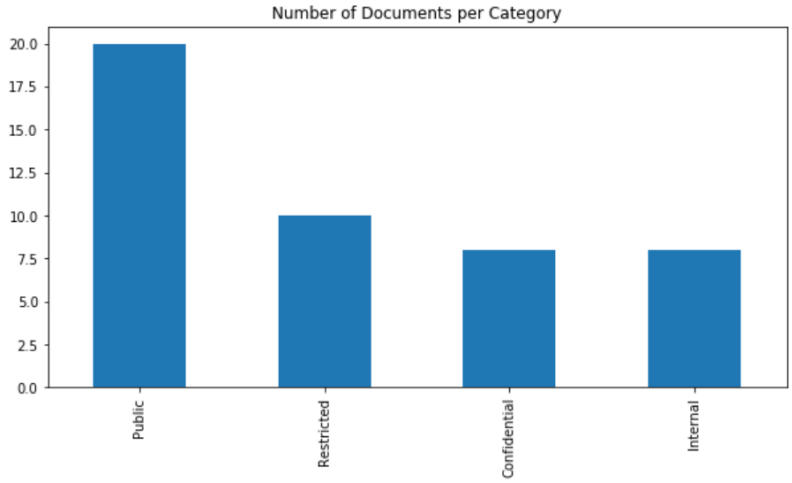


Figure No. 7.1: Bar Chart – Number of Documents per Category

Figure No. 7.2 shows the distribution of the document length (number of words).
The number of words in the document sample varies from 20 up to 37000.

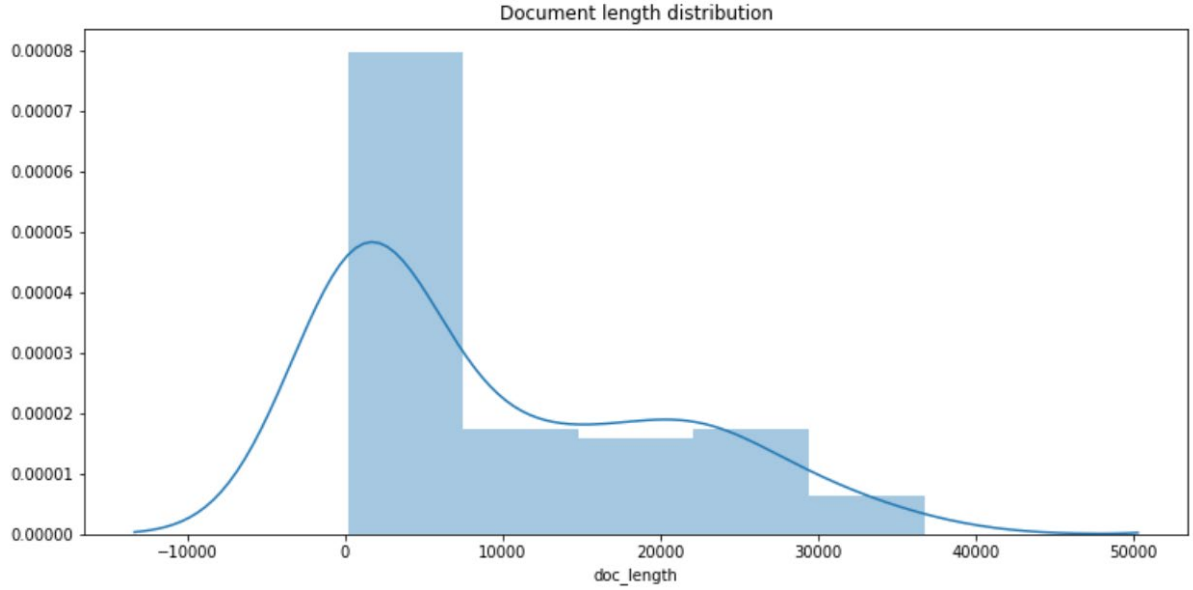


Figure No. 7.2: Document length distribution

Figure No. 7.3 shows distribution of the document length (number of words) across the different categories.

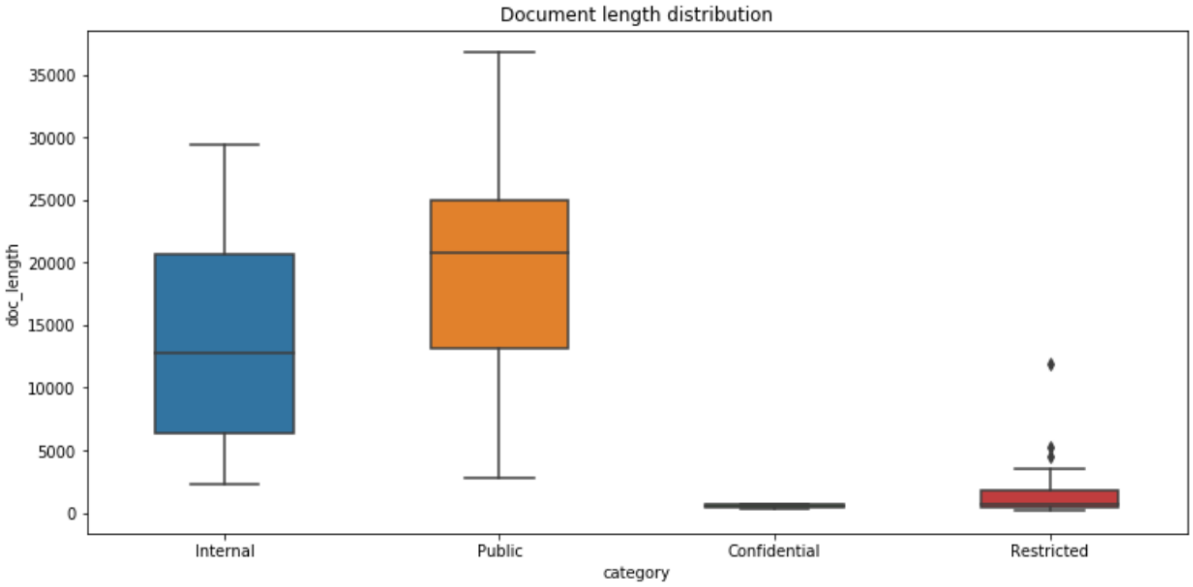


Figure No. 7.3: Document length distribution across categories

Chapter 8: Data Preparation

8.1 Synthetic Data Generation

Due to restrictions, sensitive information (PII) is unavailable on open sources.
Since the project simulation needs sensitive information, synthetic data which mimics PII was generated while preserving the format and data type. This dataset will be used as an input to train the model to auto-detect the sensitive data elements and classify the document.

Figure No. 8.1 shows a sample of how a document containing sensitive information was generated. This document contains sensitive data elements with name, DOB, email, and phone numbers.

Figure No. 8.1: Synthetic Data Generation

a) Name Generation

Names is a python library that generates random names. Figure No. 8.2 and Figure No. 8.3 shows how this library was used for generating full names of different genders.

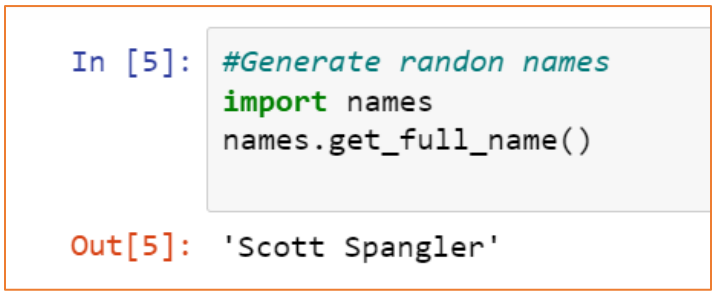


Figure No. 8.2 Name generation

```
In [11]: names.get_full_name(gender='female')
Out[11]: 'Lisa Thomas'

In [14]: names.get_full_name(gender='male')
Out[14]: 'Robert Hernandez'
```

Figure No. 8.3 Name generation by gender

b) E-mail Address generation

The names generated randomly in the previous step is used for e-mail address generation. Figure No. 8.4 shows how e-mail address are generated.

```
In [149]: email = ".".join( Names_list[0].split())+"@gmail.com"
          email
Out[149]: 'Gregory.Mccullough@gmail.com'
```

Figure No. 8.4 E-mail address generation

c) Date of birth generation

Figure No. 8.5 shows how date of births are generated. Date of births are generated randomly by incrementing from a starting date.

```
In [81]: #Generate DOB
         import numpy as np
         monthly_days = np.arange(0, 30)
         base_date = np.datetime64('1982-12-01')
         i=0
         while i <= 10:
             #random_date = base_date + np.random.choice(monthly_days)
             DOB_list.append(base_date + np.random.choice(monthly_days))
             i += 1
```

Figure No. 8.5 Date of birth generation

d) Phone Number generation

Python library random.randint(a, b) returns a random integer N such that a <= N <= b.

Figure No. 8.6 shows how Phone Number are generated using library – random.

```
In [98]: #Generate Phone Numbers
         import random
         Number_list = []
         i=0
         while i <= 100:
             Number_list.append(str(random.randint(7000000000,9999999999)))
             i += 1
```

Figure No. 8.6 Phone Number generation

e) SSN generation

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. Figure No. 8.7 shows how SSN is generated using Python library random, a set of two and four digit number is appended with the area code to generate a SSN.

```
In [121]: import names
          Names_list = []
          i=0
          while i <= 15:
              Names_list.append(names.get_full_name())
              i += 1

In [122]: #Generate SSN
          import random
          Number1_list = []
          Number2_list = []
          i=0
          while i <= 15:
              Number1_list.append(str(random.randint(10,99)))
              Number2_list.append(str(random.randint(1000,9999)))
              i += 1

In [123]: Number3_list = ['-'.join(x) for x in zip(Number1_list, Number2_list)]

In [124]: append_str = '040-'
          ssn = [append_str + sub for sub in Number3_list]
```

Figure No. 8.7 SSN generation

Each of the above generated sensitive element lists are merged into a pandas data frame and then exported as a word document.

8.2 Text Pre-processing

The unstructured word documents need to go through the process of cleaning and pre-processing to make it ready for analysis. Figure No. 8.5 shows the Text Pre-processing/ Cleaning Pipeline used for pre-processing the documents.

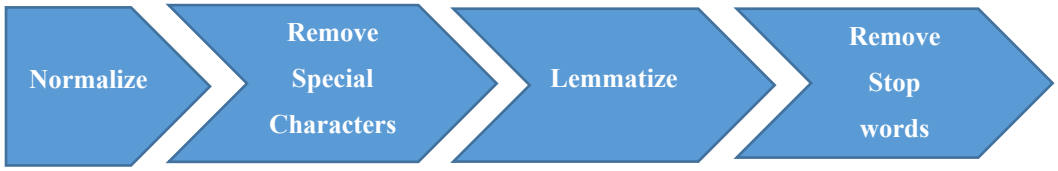


Figure No. 8.8: Text Cleaning Pipeline

Standardize

First the text is standardized by converting to lowercase.

Remove Special Characters and Punctuation

All the special characters like b', \\n, \n, \r, \\t, \\xe2, \\x80, \\x9, \\x9d, \\x94, \\x99s, \\x93, \\xc2, \\xa0, \\xef, \\x83, \\xa3, \\x82, \\xb7 are removed.

The Punctuation - !"#$%&\'()*+,-/:;<=>[\\]^_`{|}~. symbols are removed.

Lemmatize/ Normalize

Lemmatization is used to return the base or dictionary form of words (lemma). In this step we transform the words into their normalized form.

Stop word removal

All the stop words are removed from the text.

8.3 Feature Engineering

To analyse a preprocessed data, the cleaned text needs to be converted into features as per the Feature Engineering Pipeline given in the Figure No. 8.3.

Under Feature Engineering, features are created from the raw text so that the the machine learning model can be trained. The steps followed are:

- Text Featurizer
- Train-test split

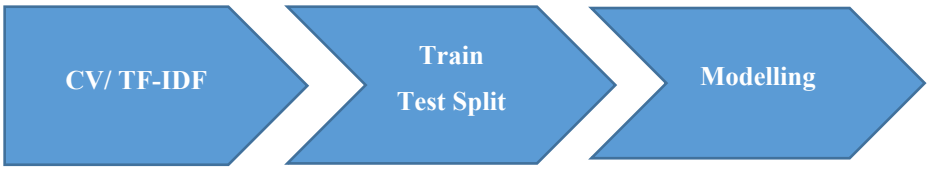


Figure No. 8.9: Feature Engineering Pipeline

Count Vectorizer

Scikit-learn's Count Vectorizer is used to convert the text documents to a vector of term/token counts. It transforms a given text into a vector on the basis of the frequency (count) of each word that occurs in the entire text.

Term Frequency – Inverse Document Frequency (TF-IDF)

TF-IDF is a weighted model which converts the text documents into vector models on the basis of occurrence of words in the documents without consideration of exact ordering.

Term Frequency (TF) – TF for a term "t" is defined as the count of a term "t" in a document "D"

Inverse Document Frequency (IDF) – IDF for a term is defined as logarithm of ratio of total documents available in the corpus and number of documents containing the term T.

To convert TF-IDF Vectors from text, the below parameters are defined:

- ngram_range: Both unigrams and bigrams are considered
- max_df: When building the vocabulary ignore terms that have a document frequency strictly higher than the given threshold.
- min_df: When building the vocabulary ignore terms that have a document frequency strictly lower than the given threshold.

- max_features: Build a vocabulary that only consider the top max_features ordered by term frequency across the corpus.

Parameter selection:

ngram_range = (1,2), min_df = 2, max_df = 1., max_features = 10000

Train Test Split

Scikit-learn's train_test_split function is used to split the dataset into training data and test data in 70-30 ratio.

Chapter 9: Data Modeling

Classification Algorithm is used to predict the class of the document. Binary classification refers to classification with only two categories. Whereas multiclass classification (multinomial classification) refers to supervised machine learning with classification of more than two classes. Since, the documents have more than one category, multi-class classification algorithm is used.

The four classes of Document Category:

1. Internal
2. Public
3. Confidential
4. Restricted

Below are the popular multi-class classification algorithms:

- K-Nearest Neighbors
- Decision Trees
- Multinomial Naïve Bayes
- Random Forest

KNN (k-nearest neighbors) Classifier

KNN or k-nearest neighbors is the simplest classification algorithm. This classification algorithm does not depend on the structure of the data. Whenever a new example is encountered, its k nearest neighbors from the training data are examined. Distance between two examples can be the euclidean distance between their feature vectors. The majority class among the k nearest neighbors is taken to be the class for the encountered example.

Decision Tree Classifier

A decision tree classifier is a systematic approach for multiclass classification. It poses a set of questions to the dataset (related to its attributes/features). The decision tree classification algorithm can be visualized on a binary tree. On the root and each of the internal nodes, a question is posed and the data on that node is further split into separate records that have different characteristics. The leaves of the tree refer to the classes in which the dataset is split.

Naive Bayes Classifier

Naive Bayes classification method is based on Bayes' theorem. It is termed as 'Naive' because it assumes independence between every pair of features in the data. Let (x1, x2, …, xn) be a feature vector and y be the class label corresponding to this feature vector.

Random Forest Classifier

The random forest is a classification algorithm consisting of many decisions trees.
It builds decision trees on different samples and takes their majority vote for classification.

After the data is cleaned and prepared, the final dataset is used for Data Modeling.
The below multi-class classification Algorithms were fit on the train data set using both Count Vectorizer and TF-IDF features.

The Table No. 9.1 illustrates the Data Modeling Train Results.

ML Classifier	Count Vectorizer	TF-IDF
	Train Accuracy	Train Accuracy
Random Forest	96.67%	98.34%
Multinomial Naïve Bayes	90%	90%
K Neighbors	73.33%	85%
Decision Tree	100%	100%

TABLE NO. 9.1: DATA MODELLING RESULTS

Chapter 10: Evaluation

10.1 Data Modeling Results

The documents were divided as train and test using the 70-30 ratio.

The Table No. 10.1 lists the modeling accuracy results:

Classifier	Count Vectorizer		TF-IDF	
	Train Accuracy	**Test Accuracy**	**Train Accuracy**	**Test Accuracy**
Random Forest	96.67%	73.08%	98.34%	80.77%
Multinomial Naïve Bayes	90%	84.62%	90%	73.08%
K Neighbors	73.33%	65.38%	85%	80.77%
Decision Tree	100%	65.38%	100%	84.62

TABLE NO. 10.1: DATA MODELLING RESULTS – TRAIN & TEST

10.2 Best Model – Multinomial Naïve Bayes using Count Vectorizer

After the model training process in the previous steps, the trained model is used to classify the document for the Test dataset.

The best model chosen for Document Classification based on train and test accuracies– **Multinomial Naïve Bayes** using a count vectorizer.

10.3 Confusion Matrix

Confusion matrix is a much better way to evaluate the performance of a classifier. To compute the confusion matrix, the set of predictions are compared against the actual targets. The idea is to count the number of times instances of class A are classified as class B. For example, the number of times the classifier confused restricted documents as confidential. A perfect classifier would have only true positives and true negatives, so it would have nonzero values only on its main diagonal.

The confusion matrix for the Multinomial Naïve Bayes algorithm is plotted to evaluate the performance of the classifier. Each rows represent the actual class, while each column represents the predicted class.

Figure No. 10.1 shows the confusion matrix for Multinomial Naïve Bayes model. It shows the kind of misclassification the algorithm makes. Some of the Restricted documents are misclassified as Confidential. Some of the Confidential documents are misclassified as Internal. Some of the Public documents are misclassified as Internal.

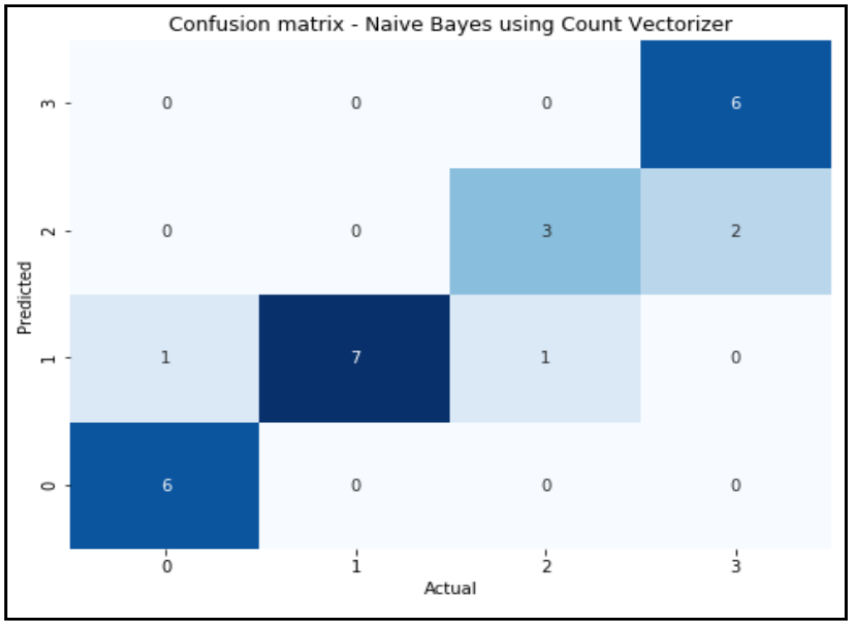


Figure No. 10.1: Multinomial Naïve Bayes – Confusion Matrix Heatmap

10.4 Classification Report Metrics

Precision

Precision of the classifier is the metric which looks at the positive predictions. It shows how accurate are the predicted positive versus the actual positive. Precision is a good measure to understand the cost of False Positive.

Precision = TP÷TP+FP (Equation No. 10.1)

Recall/TPR (True Positivity Rate)

Precision is typically used with another metric called Recall or the True Positive Rate. It is a good measure to understand the cost of False Negatives.

Recall = TP÷TP+FN (Equation No. 10.2)

F1 Score

It is often convenient to combine precision and recall into a single metric called the F1 score. It is the harmonic mean of precision and recall. High F1 score means that both Precision and Recall are high.

F1 score = 2×(Precision×Recall)÷(Precision+Recall) (Equation No. 10.3)

TP – The number of True Positives

FP – The number of False Positives

FN – The number of False Negatives

Based on the train and test accuracies, Multinomial Naïve Bayes classification model was chosen as the best model. The Figure No. 10.2 shows the Classification Report – Precision, Recall and F1 score for this chosen classification model.

Classification report

	precision	recall	f1-score	support
1	0.86	1.00	0.92	6
2	1.00	0.78	0.88	9
3	0.75	0.60	0.67	5
4	0.75	1.00	0.86	6
accuracy			0.85	26
macro avg	0.84	0.84	0.83	26
weighted avg	0.86	0.85	0.84	26

Figure No. 10.2: Multinomial Naïve Bayes – Classification Report

Chapter 11: Deployment

This python model can be easily integrated with any data cataloguing and data discovery framework where there is no capability of scanning unstructured data sources. This can be customized according to privacy and classification needs and deployed on on-premise and cloud infrastructure platforms via APIs.

Figure No. 11.1 shows the deployment model which can be deployed using the Flask server. It can then be used to scan unstructured documents from an organization's SharePoint, Confluence, etc.
Results from the scanning of unstructured documents:

1. Identify sensitive data elements using Regex.

2. Classify the document from the best model that was saved.

3. Score the document based on 1 and 3 above.

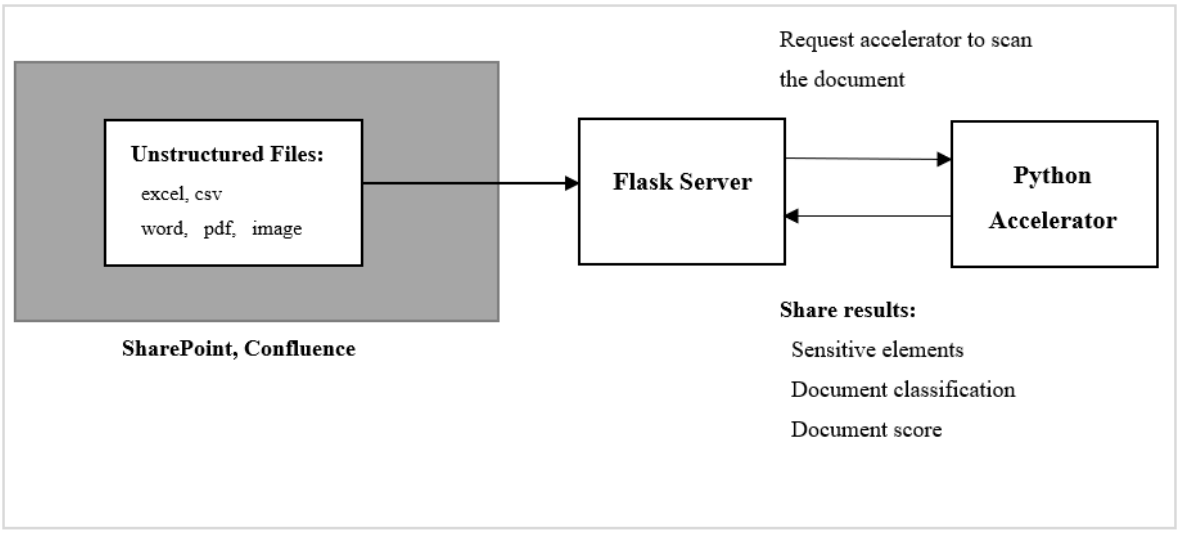


Figure No. 11.1: Deployment Model

This python model can be used by an organization to build their classifiers, custom attributes, and meta-models to accurately identify and classify sensitive data across all the unstructured data sources.

Chapter 12: Analysis and Results

Part 1 Results – Identifying Sensitive Data Elements

a) Person's Name

Figure No. 12.1.1 shows an example of identifying a name contained in the document as per the format, pattern, and keyword discussed in Chapter 5.1.1. The figure highlights the keyword 'name' in the document along with 11 names.

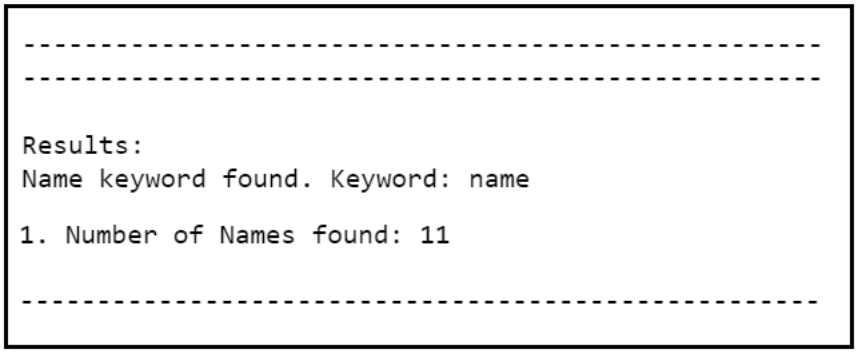


Figure No. 12.1.1: Results – Person's Name

b) Phone Numbers

Figure No. 12.1.2 shows an example of identifying phone numbers in the document as per the format, pattern, and keyword discussed in Chapter 5.1.4. The figure highlights the keyword 'number' in the document along with the number of phone numbers found.

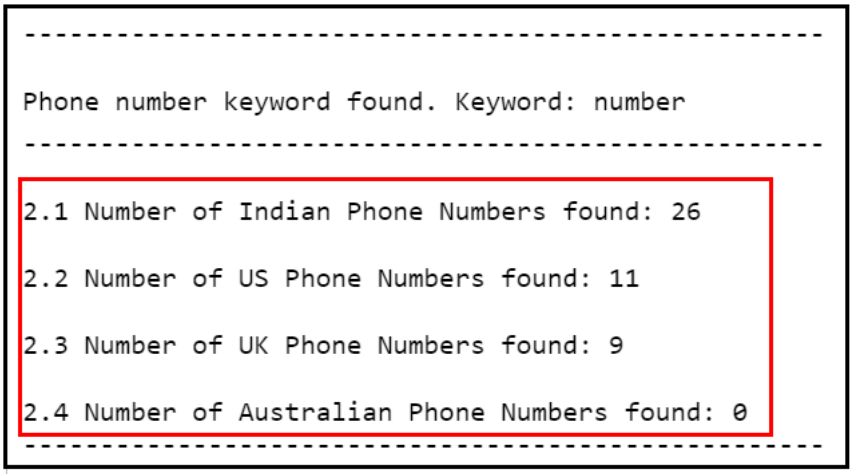


Figure No. 12.1.2: Results – Phone Number

c) Email Address

Figure No. 12.1.3 shows an example of identifying email address in the document as per the format, pattern, and keyword discussed in Chapter 5.1.2. The figure highlights the keyword 'email' found in the document along with the number of email address found.

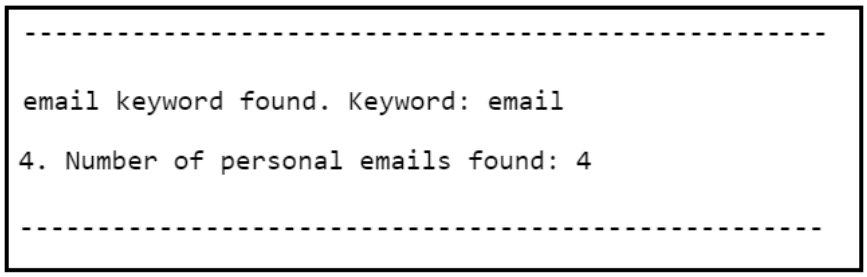


Figure No. 12.1.3: Results – email address

d) SSN/ PAN

Figure No. 12.1.4 shows an example of identifying PAN/SSN in the document as per the format, pattern, and keyword discussed in Chapter 5.1.5 and Chapter 5.1.6. The figure highlights the keyword 'PAN/SSN' found in the document along with the number of PAN and SSN numbers found in the document.

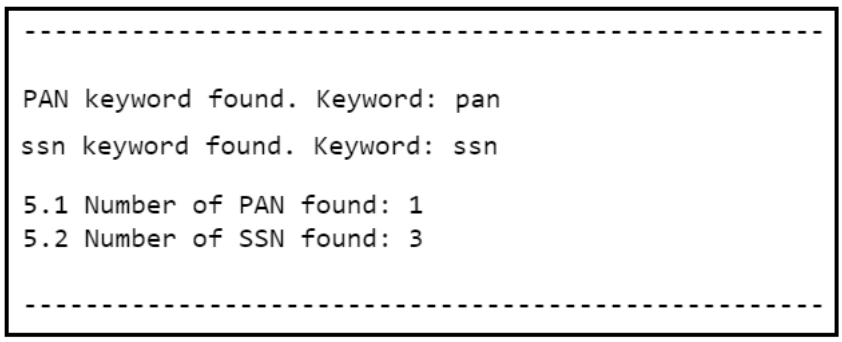


Figure No. 12.1.4: Results – SSN/PAN

Figure No. 12.1.5 shows an example of document classification and document risk categorization.

```
Calculating the risk score....
--------------------------------------------------
--------------------------------------------------

Summary:

The Data Security Classficiation of the document is: RESTRCITED
The Document Risk Categorization for the document is: High Risk

--------------------------------------------------
End
```

Figure No. 12.1.5: Results – Data Security Classification

Part 2 Results – Document Classification

The document was cleaned and processed. The document was classified as a Restricted Document using the naïve bayes model trained in chapter 9.

Part 3 Results – Risk Categorization

Table No. 12.1 shows the Risk Categorization defined in Chapter 5.3. The document is categorized as a High-Risk document since it contains restricted elements – SSN and PAN along with names and phone numbers.

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

TABLE NO. 12.1 : DOCUMENT RISK CATEGORIZATION

Part 4 Results – Executive Dashboard

Figure 12.2 shows the Executive Tableau Dashboard built to summarize the results of the documents scanned. It highlights the number of documents reviewed, Document Category and Risk and the number of sensitive elements found.

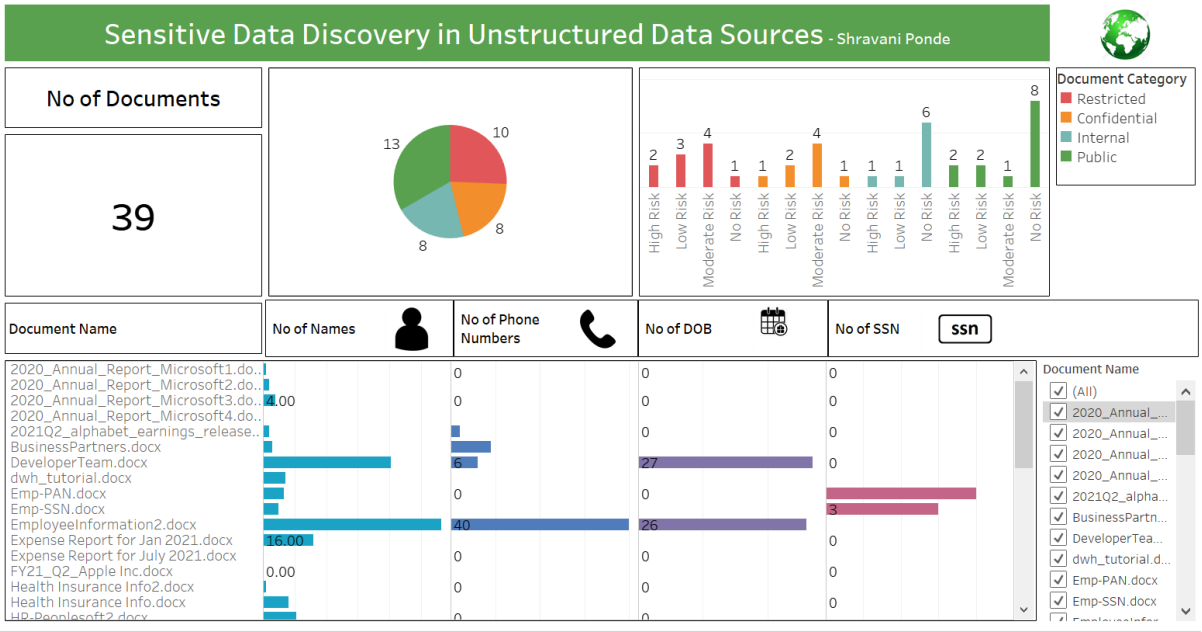


Figure No. 12.2: Executive Dashboard

Chapter 13: Conclusions and Future Scope

13.1 Conclusion

The discovery results and the risk scores are produced as a Dashboard, which allows the business stakeholders to take necessary actions in protecting their sensitive data assets from heterogeneous unstructured sources.

13.2 Future Scope

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

Enabling Data activity monitoring controls

Data Protection capabilities

- Masking
- Tokenization
- Anonymization
- Pseudonymization
- Encryption
- Data Loss Preventions (DLP)
- Data Remediation
- Data Subject Rights

Bibliography

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

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

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

Gai, K., Qiu, M., & Zhao, H. (2017). Privacy-Preserving Data Encryption Strategy for Big Data in Mobile Cloud Computing. *IEEE*, 678 - 688.

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

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

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

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

Marr, B. (2018, May 21). *How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read*. Retrieved from Forbes: 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. (2016). Protection of Big Data Privacy. *IEEE*, 1821 - 1834.

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

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

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

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

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

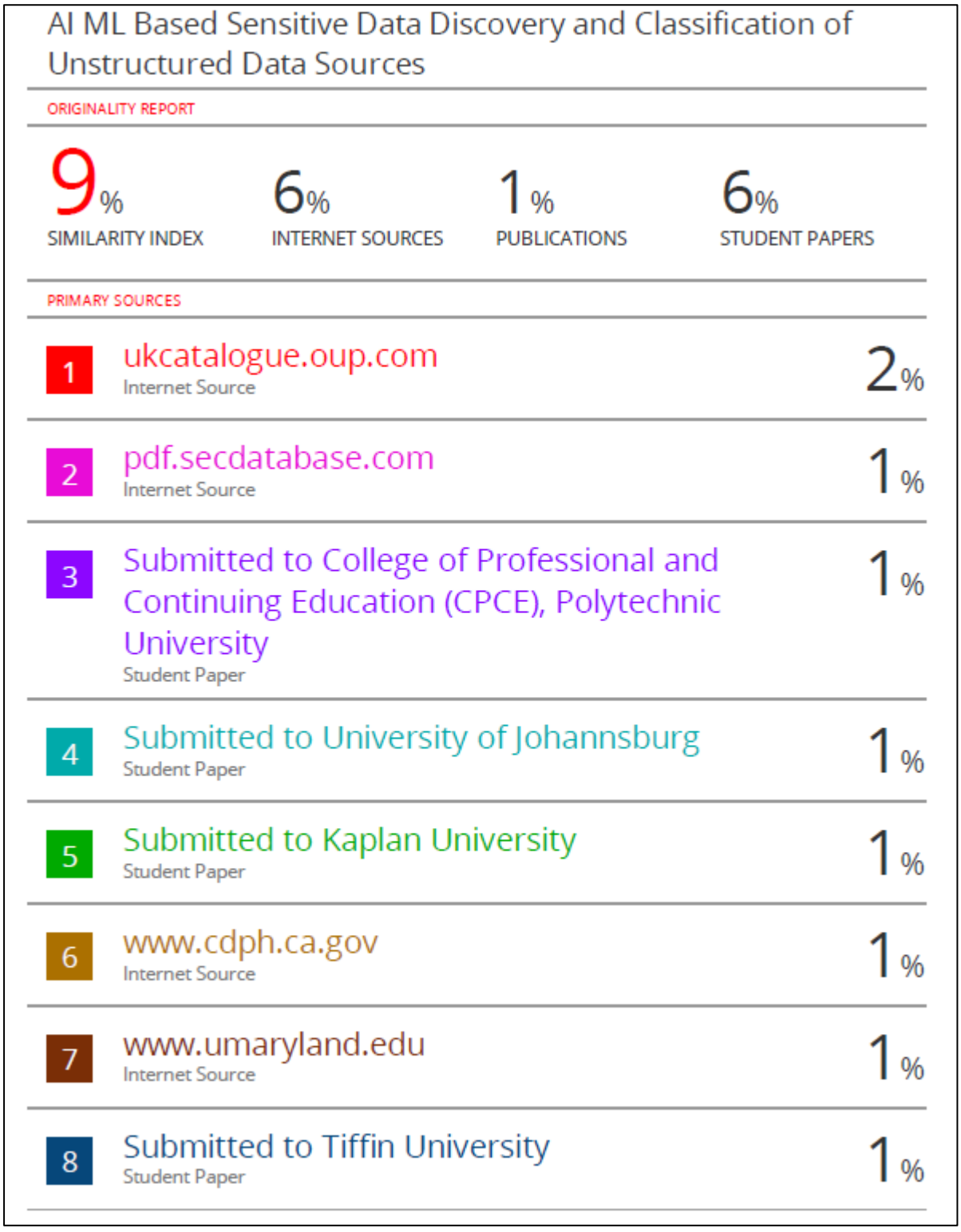
Yaqoob, I., Salah, K., Jayaraman, R., & Al-Hammadi, Y. (2022). Blockchain for healthcare data management: opportunities, challenges, and future recommendations. *Springer Link*, 11475–11490.

Yeh, S. -C.-H. (2018). A Data-Driven Security Risk Assessment Scheme for Personal Data Protection. *IEEE Access, vol. 6, pp. 50510-50517*.

Zhang, X., Qi, L., Dou, W., He, Q., Leckie, C., Kotagiri, R., & Salcic, Z. (2017). MRMondrian: Scalable Multidimensional Anonymisation for Big Data Privacy Preservation. *IEEE*, 125 - 139.

Appendix

Plagiarism Report[1]



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

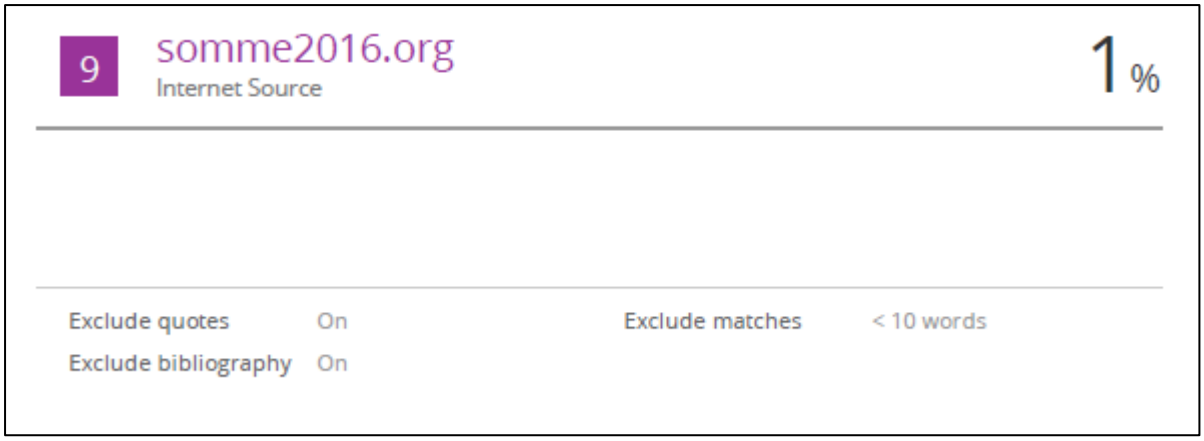
ORIGINALITY REPORT

9%	6%	1%	6%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	ukcatalogue.oup.com Internet Source	2%
2	pdf.secdatabase.com Internet Source	1%
3	Submitted to College of Professional and Continuing Education (CPCE), Polytechnic University Student Paper	1%
4	Submitted to University of Johannsburg Student Paper	1%
5	Submitted to Kaplan University Student Paper	1%
6	www.cdph.ca.gov Internet Source	1%
7	www.umaryland.edu Internet Source	1%
8	Submitted to Tiffin University Student Paper	1%

[1] Turnitn report to be attached from the University.


9
somme2016.org
Internet Source
1 %
Exclude quotes On
Exclude bibliography On
Exclude matches < 10 words

Publications in a Journal/Conference Presented/White Paper[2]

Author's Name – Shravani Ponde

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

Conference Name - EAI ICISML 2022

Submitted Date - 20-10-2022.

Paper ID - 324099

Conference scheduled for - December 2022.

[2] URL of the white paper/Paper published in a Journal/Paper presented in a Conference/Certificates to be provided.

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

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.

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 & 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.

Fig. 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

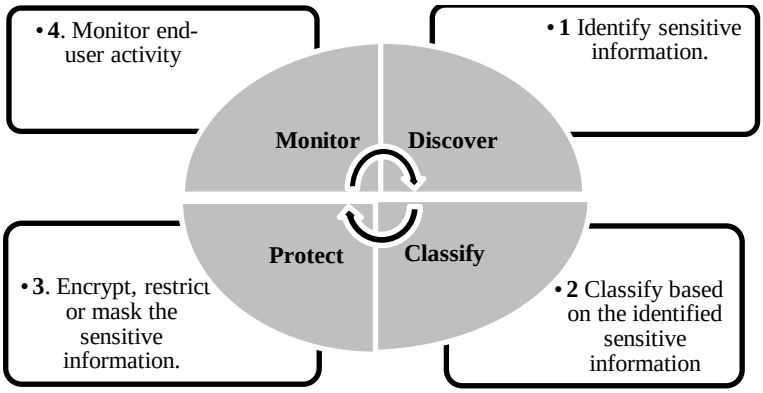


Fig. 1. Data Protection Lifecycle

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.

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 III. 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		
	Pattern	**Sample**
SSN	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)

Keywords: ssn, social security number, ssn number, social security #, social security no, soc sec, ssn#

E-mail address

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

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

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

The format and pattern for other sensitive elements like Name, Phone number, Date of birth etc., was defined similarly.

4.2 Rule based 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. Fig. 2. shows the text cleaning pipeline used for pre-processing the documents.

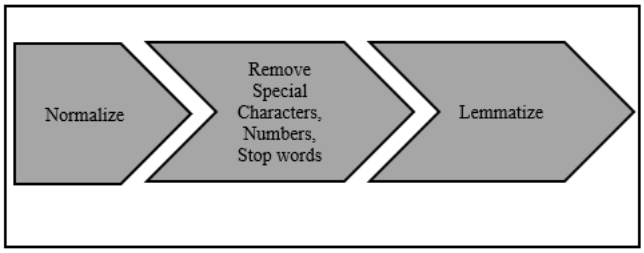


Fig. 2. Text pre-processing pipeline

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.

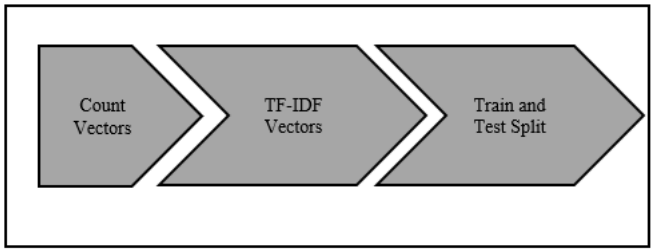


Fig. 3. Feature engineering pipeline

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

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

David, D. (2019, July 9). *AI Unleashes the Power of Unstructured Data*. Retrieved from CIO: 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. (2020, December 14). *The Rising Concern Around Consumer Data And Privacy*. Retrieved from Forbes: https://www.forbes.com/sites/forbestechcouncil/2020/12/14/the-rising-concern-around-consumer-data-and-privacy/?sh=30741b43487e

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

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

Marr, B. (2018, May 21). *How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read*. Retrieved from Forbes: 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. (2016). Protection of Big Data Privacy. *IEEE*, 1821 - 1834.

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

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

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

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

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

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

Zhang, X., Qi, L., Dou, W., He, Q., Leckie, C., Kotagiri, R., & Salcic, Z. (2017). MRMondrian: Scalable Multidimensional Anonymisation for Big Data Privacy Preservation. *IEEE*, 125 - 139.