Personally Identifiable Information Analysis

A Project Report

Submitted by:

Piyush Kumar (22030142020)

in partial fulfillment for the award of the degree

of

M.SC. COMPUTER APPLICATION

IN

COMPUTER STUDIES

at

SYMBIOSIS INSTITUTE OF COMPUTER STUDIES AND RESEARCH, PUNE, INDIA

AFFILIATED TO SYMBIOSIS INTERNATIONAL (DEEMED UNIVERSITY) (INDIA)

September 2023

DECLARATION

I hereby declare that the project entitled “Personally Identifiable Information Analysis” submitted for the M.Sc. (Computer Application) degree is my original work and the project has not formed the basis for the award of any other degree, diploma, fellowship or any other similar titles.

Piyush

Signature of the Student

Place: PUNE

Date: 01/09/2023

CERTIFICATE

This is to certify that the Project Pilot titled "Personally Identifiable Information Analysis" is the bona fide work carried out by Piyush Kumar (22030142020), a student of M.Sc. Computer Application of Symbiosis Institute of Computer Studies and Research, Pune India, affiliated to Symbiosis International (Deemed University) during the academic year 2022-24, in partial fulfillment of the requirements for the award of the degree of M.Sc. Computer Application.

SJoshi

Signature of the Guide

Place: PUNE

Date: 01/09/2023

Acknowledgement

We would like to thank Mihir Vohra and Shirish Joshi for their essential contributions to our pilot study on the investigation of Personally Identifiable Information (PII). Their active participation was critical in acquiring critical information and varied perspectives for our study, which delves into the complex environment of privacy and security issues linked with personal data.

Abstract

Purpose: -

The purpose of this study is to look into the risks associated with exposing personally identifiable information (PII). As digital technology and online activities become more ubiquitous, the risk of PII disclosure has grown to be a serious worry for individuals, businesses, and regulators. This study intends to identify the essential elements that enhance the risks associated with PII disclosure through the use of risk mitigation measures.

Methodology: -

A thorough literature review was performed to learn more about PII and its analysis. The study looked at a variety of PII types, including name, emails, phone number, and social security number. The study also looked into how PII is collected, including online forms, social media, and public records. The study also looked at the dangers of identity theft, financial fraud, and illegal access that could emerge from the misuse of PII.

Major Findings: -

The study's findings found that, while there is a general grasp of the need of preserving PII, there is a lack of awareness among employees regarding specific data protection rules and best practices. Furthermore, some firms are not establishing proper safeguards for PII, which can result in data breaches and associated legal ramifications. According to the report, firms should focus employee training and education on PII protection and data security, as well as implementing effective measures such as data encryption and access controls.

Table of Contents

1. Introduction: -

1.1 Problem Definition: -

A person's name, social security number, email address, phone number, and other personally identifiable information (PII) are all examples of information that can be used to identify them. As more and more data are collected and kept digitally, it is becoming increasingly important to protect this sensitive information from data breaches and misuse. PII analysis refers to the process of recognizing and evaluating potential threats to PII in order to protect its confidentiality, integrity, and availability.

1.2 Problem Overview: -

PII stands for Personally Identifiable Information. It is any piece of information that can be used to identify a specific person, whether used alone or in conjunction with other pieces of information.

The problem with PII is that it is sensitive data that can be utilized for fraud, identity theft, and other illegal activities. It is critical to safeguard information from unauthorized access, use, disclosure, and destruction.

To address this issue, the management of PII must be explicitly specified. Guidelines and practices for gathering and storing should be included in these standards.

1.3 Hardware Specification: -

Specific hardware requirements will be required depending on the size and complexity of the dataset being investigated, the analytical tools and algorithms being used, and the desired performance levels.

A CPU with multiple cores and rapid clock speeds is often recommended for PII investigation. It is best to get an Intel or AMD CPU with four or more cores, such as the Intel Core i7 or AMD Ryzen 7. Data analysis necessitates the use of random-access memory (RAM), and the bulk of PII analysis tasks necessitate at least 8GB of RAM. However, larger and more complex datasets may necessitate 16GB of RAM or more. In addition to the CPU and memory, storage is an important consideration.

1.4 Software Specification: -

The software that handles Personally Identifiable Information (PII) must be built to comply with data protection requirements such as GDPR and CCPA. It should have strong authentication, data encryption, regulated data access, disclosure monitoring, and appropriate PII disclosure request protocols. Data retention policies that ensure data is stored just as long as necessary are critical. To avoid breaches and protect PII, the software's dependability, scalability, speed, and adherence to data security best practices are critical. For development, a Python IDE and the necessary libraries are suggested.

2. Literature Survey: -

2.1 Existing System: -

The following systems are typically used to detect PII information:

"Data loss prevention" (DLP) systems are systems that identify and prevent the loss of sensitive data. They can be configured to track a range of channels, such as email, web traffic, and file transfers, and to detect PII data based on pre-defined rules.

Machine learning (ML) models can be trained to recognize patterns and find personally identifying information. These models, which may be trained using labelled data, can be used to find personally identifiable information (PII) in text or images.

2.2 Proposed System: -

- Obtain user consent: Before scanning a user's system or files for Personally Identifiable Information (PII), the user's explicit permission will be obtained.
- Scan for PII: Using high-level computer languages and open-source tools or available GitHub repositories, the user's system or files will be checked for specified PII, such as email addresses and emails.
- Display PII score: A PII score will be calculated and displayed to the user in a straightforward format based on the identified quantity and categories of PII.
- Make recommendations: Based on the PII score, practical recommendations will be made to improve system security. High scores (above 20) may elicit solutions from Galaxkey Enterprises, while low scores (20 or below) may prompt optional steps for improvement.
- Overall, our goal is to assist customers in recognizing security issues and providing practical recommendations on how to protect their sensitive information.

2.3 Feasibility Study: -

The purpose of this feasibility study is to determine the viability and feasibility of developing a detection and encryption system for the project to analyze personally identifiable information. After scanning the system for a PII score, the technology will be designed to automatically encrypt or conceal critical information to prevent unauthorized access.

The recommended solution is technically feasible because encryption and detection are both covered by current technology. Galaxkey Enterprises, Cipher Cloud, and Secure Age Technologies offer solutions that might be incorporated into the proposed system.

Financial viability: The development and deployment of the proposed system will necessitate significant financial resources. Hardware, software, and human resource expenses must all be considered. A thorough financial analysis will be required.

Operational Suitability: The proposed system will require skilled employees to operate and maintain it. Staff will require enough training in the required technology and security procedures. To ensure that the system is used to its full potential, an operating strategy and training program must be developed.

2.4Future Scope: -

- Data privacy and security may benefit greatly in the future from the analysis of unstructured data, such as audio, video, photos, text files, etc. The possibility of collecting private and sensitive data grows as the usage of IoT devices expands. As a result, it is imperative to put in place efficient safeguards for user privacy.
- Using PII detection software at the network layer to remove sensitive data and only deliver pertinent data for analytics is one possible option. Using this strategy, it will be less likely that private data will be sent to outsiders or kept in dangerous places.
- When PII is shouted aloud when dealing with IoT devices, another possibility is to disable the microphone at the software level. This may be done by identifying when confidential information is being uttered and stopping the device from transmitting it using speech recognition software or other comparable technologies.
- A further layer of defence against the unlawful gathering and use of personal information may be added by integrating PII detection technologies into various apps as an accessible API or SDK.

- Developers may simply include the capability of the PII detection technology into their apps by making it available as an API or SDK. This allows developers to quickly and easily discover and secure any sensitive data that their applications may be managing. For instance, an app developer may check user-entered data for sensitive information like social security numbers, credit card numbers, or other personal information that has to be secured. This could be done using a PII detection SDK.
- Overall, in the era of IoT devices, it is critical to take into account and adopt a variety of solutions to secure sensitive information. To guarantee that user privacy is preserved, this may entail a mix of technological, legal, and regulatory safeguards.
- Exposing PII detection as an API or SDK can assist to increase user confidence in the application by demonstrating that the developer takes privacy and security seriously in addition to safeguarding users' sensitive information.
- In general, making PII detection technologies available as an API or SDK can be a good method to ensure that personal data is kept private and safe and to give consumers more confidence when using apps that handle their sensitive data.

2.5 Industry Requirements: -

- Several sectors place a high priority on protecting personally identifiable information (PII), and there are particular rules and requirements that businesses must follow to protect the security and privacy of this data. The following are a few industry standards for the analysis of PII information:
- Healthcare: The Health Insurance Portability and Accountability Act (HIPAA), which establishes stringent guidelines for the security of patient information, applies to the healthcare sector. All PII, such as patient names, addresses, and medical information, must be encrypted both in transit and at rest, according to organisations. Furthermore, to restrict who may read or alter PII, healthcare institutions must have thorough access controls in place.
- Education: The protection of student information in the education sector is governed by the Family Educational Rights and Privacy Act (FERPA). Names, addresses, and academic records of students are examples of PII in this category. Educational institutions must make sure that this data is maintained securely and that only authorised individuals have access to it.
- We have learned from https://www.galaxkey.com/ that they are providing a desktop programme for PII information analysis. According to the description, their solution appears to be intended to assist enterprises in safeguarding their personally identifiable information

(PII) by transferring it to a secure area within their system and then encrypting that location to deter illegal access.

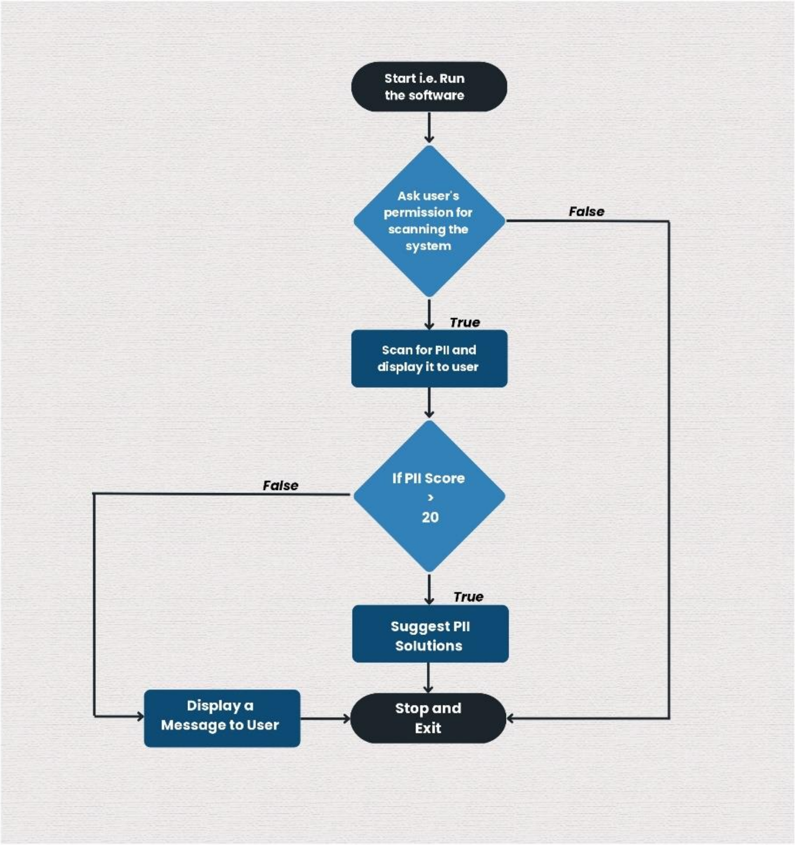
3. System Analysis and Design: -

3.1 User Requirement Specifications: -

The following are the user requirements for the PII analysis: -

- The user must give permission for the system to search the entire system for PII data.
- The system will calculate a PII score based on the scans.
- The user will receive an easy-to-understand graphical representation of the PII score.
- Based on the PII score, the algorithm will provide the best strategies for mitigating the privacy concerns associated with PII data.
- The system will adhere to moral norms for investigation and analysis.

3.2 Flowchart: -


Start i.e. Run the software
Ask user's permission for scanning the system
False
True
Scan for PII and display it to user
If PII Score > 20
False
True
Suggest PII Solutions
Display a Message to User
Stop and Exit

3.3 Algorithm:

- Import Libraries:
    - Import the required libraries for regular expressions, data visualization using Plotly and Matplotlib.
- Define PII Patterns and Weights:
    - Create a dictionary named patterns that associates PII types with their corresponding regex patterns and weights.
- PII Detection Function:
    - Define a function named detect_pii(text) that takes a text input and returns a dictionary of detected PII types along with matched instances.
    - Loop through each PII type in the patterns dictionary.
    - Use regex to find all matches of the PII pattern in the given text.
    - If matches are found, add the PII type and matched instances to the dictionary.

- Display Results Function:
    - Define a function named display_results(scanned_text, detected_pii, pii_score) to visualize the PII detection results and PII score.
    - Calculate the counts of detected PII and total PII occurrences for each PII type.
    - Create a bar chart using Matplotlib to compare detected PII counts with total PII counts.
    - Create a gauge indicator using Plotly to display the PII score.
- Scan System for PII Function:
    - Define a function named scan_system_for_pii(file_path) that reads text from a given file path and returns the text content.
- User Consent and Processing:
    - Ask the user for consent to scan the system for PII.
    - If the user agrees, proceed with the following steps:
    - Specify the file path containing the text to be scanned.
    - Use the scan_system_for_pii function to read the text from the file.
    - Use the detect_pii function to detect PII in the text.
    - Calculate the PII score by considering the weighted counts of detected PII instances.
    - Call the display_results function to visualize the PII detection results and PII score.
- Based on the PII score:
    - If the score is high (greater than a threshold), provide recommendations for improving system security.
    - If the score is low, provide suggestions for enhancing security.
- User Consent Denial:
    - If the user does not provide consent, display a message indicating that PII scanning requires user consent.

3.4 Testing Process: -

- **Testing at the Unit Level:**
    - The detect_pii function was tested for accuracy using a range of inputs, including several sorts of PII patterns.
- **Testing for Integration: Validate the entire process: -**
    - Scan a test file containing known PII to check that the identified PII matches the actual information.

    - Manually compute the PII score based on patterns and weights, then compare it to the shown value.

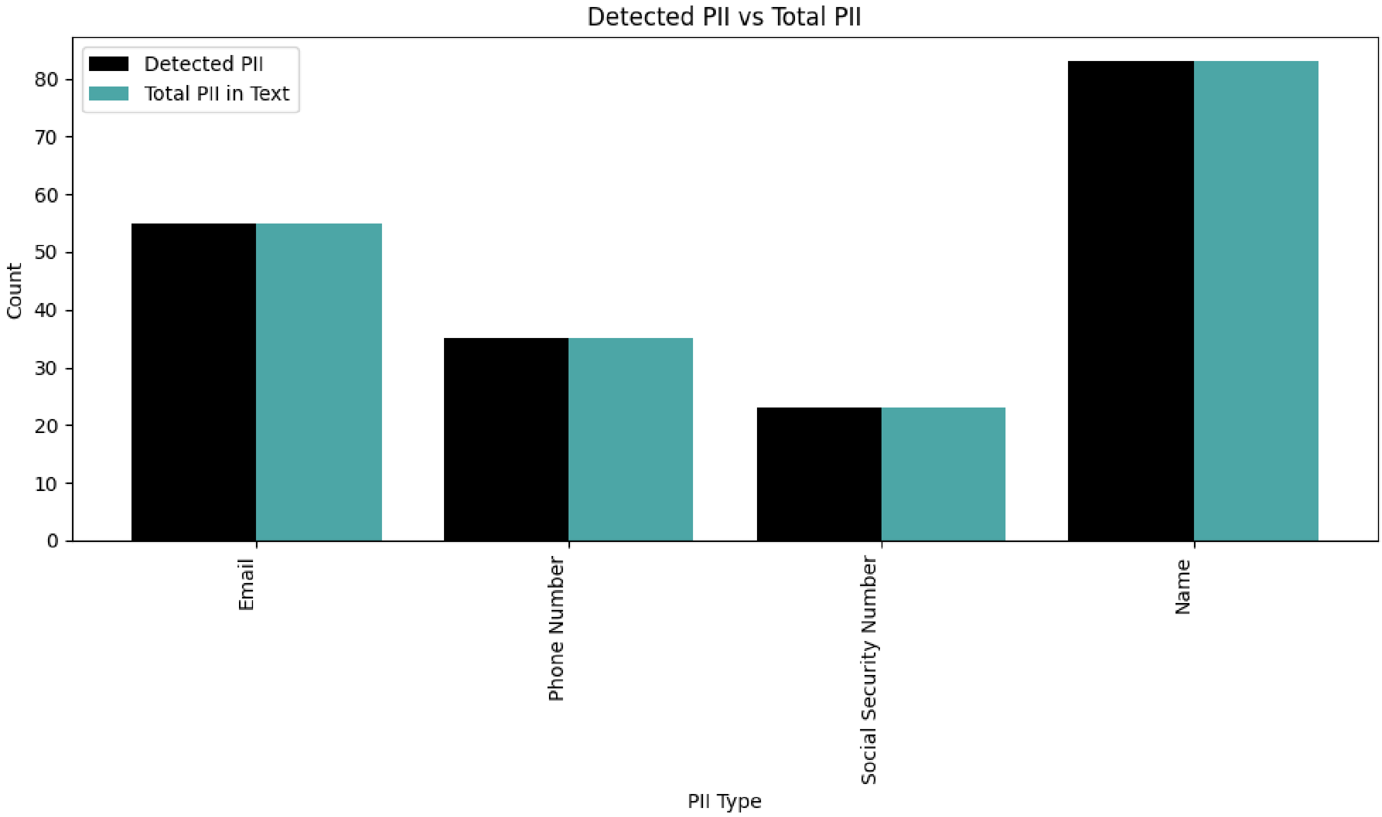
- **Visualizations Testing:**
    - Investigated the bar chart visualization.
    - Mock data was used to depict various scenarios.
    - It was verified that the detected and total PII counts for each kind were accurate.
    - Ensured the accuracy of labels, colors, look, and layout.
    - The gauge indicator visualization was evaluated.
    - Input numerous PII scores and ensure that the gauge displayed the correct values.
    - Checked to see if the gauge changed in response to the PII score.

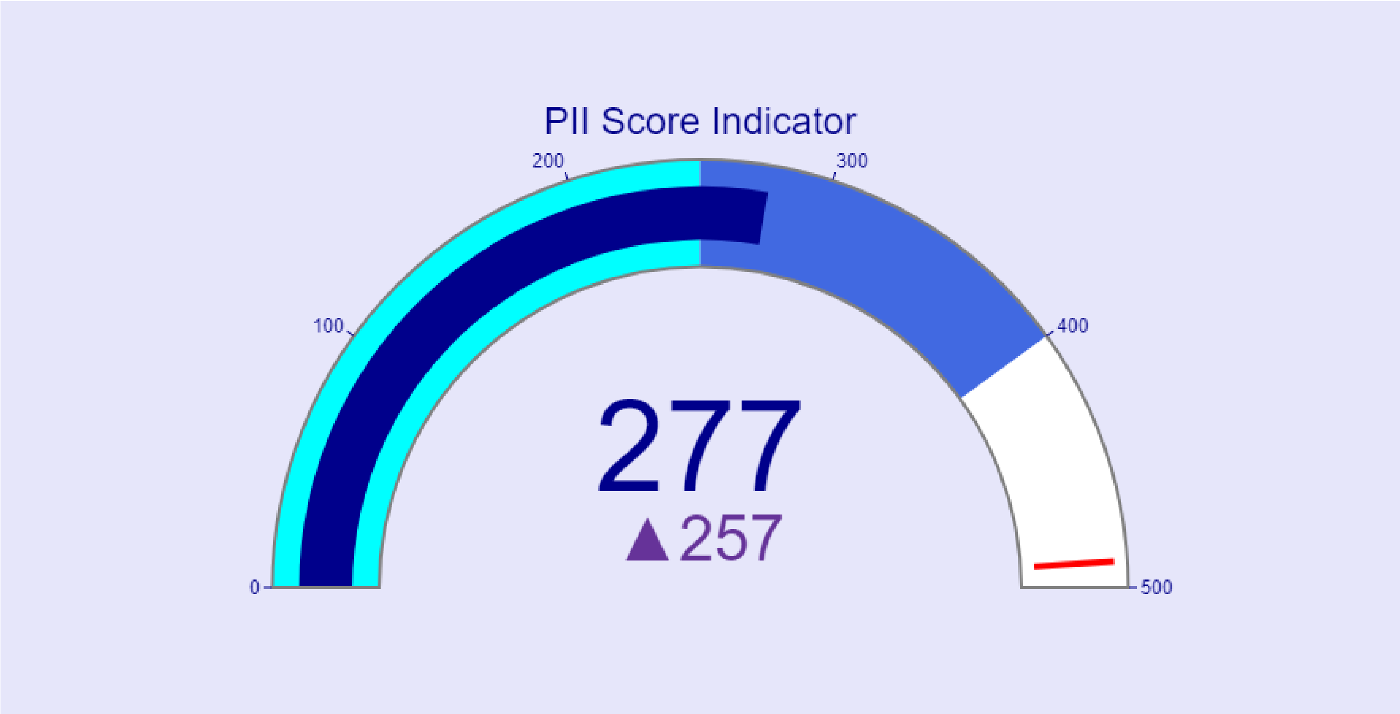
3.5 Code Execution and Output: -

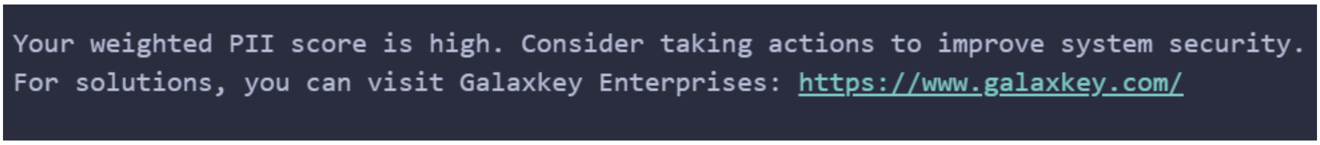
```
import re  # Importing the regular expressions library.
import plotly.graph_objects as go  # Importing Plotly for interactive data visualization.
import matplotlib.pyplot as plt  # Importing Matplotlib for static data visualization.

# Define patterns for different types of PII information along with weights.
patterns = {
    "Email": {"pattern": r"\b[A-Za-z0-9._%+-]+@[A-Za-z0-9.-]+\.[A-Z|a-z]{2,}\b", "weight":
1},
    "Phone Number": {"pattern": r"\b(?:\+?1[-.\s]?)?\(?\d{3}\)?[-.\s]?\d{3}[-
.\s]?\d{4}\b", "weight": 2},
    "Social Security Number": {"pattern": r"\b\d{3}[-.\s]?\d{2}[-.\s]?\d{4}\b", "weight":
3},
    "Name": {"pattern": r"\b[A-Z][a-z]+\s[A-Z][a-z]+\b", "weight": 1}
}

# Function to find PII information in text using patterns.
def detect_pii(text):
    detected_pii = {}
    for key, pattern_info in patterns.items():
        matches = re.findall(pattern_info["pattern"], text)
        if matches:
            detected_pii[key] = matches
    return detected_pii

# Function to display the PII score and Bar chart.
def display_results(scanned_text, detected_pii, pii_score):
    # Calculate the count of Detected and Total PII occurrences.
    detected_pii_counts = {key: len(matches) for key, matches in detected_pii.items()}
    total_pii_counts = {key: len(re.findall(pattern_info["pattern"], scanned_text)) for
key, pattern_info in patterns.items()}
```

```
    # Create a Bar chart to visualize Detected PII vs Total PII.
    plt.figure(figsize=(10, 6))
    bar_width = 0.4
    index = range(len(detected_pii_counts))
    plt.bar(index, detected_pii_counts.values(), bar_width, color='black', label='Detected 
PII')
    plt.bar([i + bar_width for i in index], total_pii_counts.values(), bar_width, 
color='teal', alpha=0.7, label='Total PII in Text')
    plt.xlabel('PII Type')
    plt.ylabel('Count')
    plt.title('Detected PII vs Total PII')
    plt.xticks([i + bar_width / 2 for i in index], detected_pii_counts.keys(), 
rotation=90, ha='right')
    plt.legend()
    plt.tight_layout()
    plt.show()

    # Create a gauge indicator to display the PII score.
    fig = go.Figure(go.Indicator(
        mode="gauge+number+delta",
        value=pii_score,
        domain={'x': [0, 1], 'y': [0, 1]},
        title={'text': "PII Score Indicator", 'font': {'size': 24}},
        delta={'reference': 20, 'increasing': {'color': "RebeccaPurple"}},
        gauge={
            'axis': {'range': [None, 500], 'tickwidth': 1, 'tickcolor': "darkblue"},
            'bar': {'color': "darkblue"},
            'bgcolor': "white",
            'borderwidth': 2,
            'bordercolor': "gray",
            'steps': [
                {'range': [0, 250], 'color': 'cyan'},
                {'range': [250, 400], 'color': 'royalblue'}],
            'threshold': {
                'line': {'color': "red", 'width': 4},
                'thickness': 0.75,
                'value': 490}}))
    fig.update_layout(paper_bgcolor="lavender", font={'color': "darkblue", 'family': 
"Arial"})
    fig.show()

# Function to read text from a file.
def scan_system_for_pii(file_path):
    with open(file_path, "r") as file:
        text_to_scan = file.read()
    return text_to_scan

# Ask user for consent before scanning for PII.
user_consent = input("Do you agree to a PII scan of your system? (yes/no): ")
if user_consent.lower() == "yes":
    file_path = "D:\\Symbiosis Projects\\Pilot_Project\\PII_Data.txt"  # Define the file
path.
    scanned_text = scan_system_for_pii(file_path)  # Read text from the file.
    detected_pii = detect_pii(scanned_text)  # Detect PII in the text.

    # Calculate the PII score by considering the weighted counts.
    pii_score = sum(len(matches) * pattern_info["weight"] for key, matches in
detected_pii.items() for pattern, pattern_info in patterns.items() if pattern == key)

    # Display the results.
    display_results(scanned_text, detected_pii, pii_score)

    # Provide recommendations based on the PII score.
    if pii_score > 20:
        print("Your weighted PII score is high. Consider taking actions to improve system
security.")
        print("For solutions, you can visit Galaxkey Enterprises:
https://www.galaxkey.com/")
    else:
        print("Your weighted PII score is low. You can still take steps to enhance system
security.")
        print("For solutions, you can visit Galaxkey Enterprises:
https://www.galaxkey.com/")
else:
    print("PII scanning requires user consent.")
```

Output: Consent – Yes

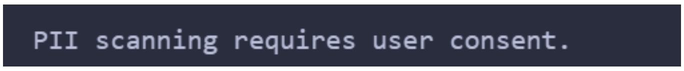






```
Your weighted PII score is high. Consider taking actions to improve system security.
For solutions, you can visit Galaxkey Enterprises: https://www.galaxkey.com/
```

Output: Consent – No



```
PII scanning requires user consent.
```

4. Encountered Challenges: -

We encountered several challenges while analyzing personally identifiable information (PII), and these challenges can be outlined as follows:

- Gathering PII-Containing Data: One of the primary difficulties we encountered was related to the sensitivity of PII. Acquiring data that included PII proved to be a challenging task. Due to the sensitive nature of this information, it was not readily available, making the data collection process arduous.
- Limited PII Data on Open Source Platforms: Many open source platforms were found lacking in PII-containing datasets, primarily due to security concerns. In cases where such data was present, it often came at a significant cost. This scarcity of accessible PII data hindered our analysis efforts.
- Complexities in Regex Pattern Matching: Another substantial challenge we faced was creating regex patterns to identify PII. PII data comes in various formats, and this lack of

uniformity posed a significant hurdle. For instance, addresses differ in format between countries like the USA and India. Additionally, unexpected variations, such as names including numbers, added another layer of complexity to the regex pattern matching process.

- Contextual Identification of PII: Identifying PII within the context of documents was an additional challenge. This context-specific aspect could potentially be addressed using machine learning techniques. However, implementing such solutions might require a considerable allocation of resources and developer expertise.

5. Conclusion: -

Individuals can be identified and possibly damaged if personally identifiable information (PII), a highly sensitive data set, falls into the wrong hands. Individuals and organizations must recognize the importance of protecting PII and take the necessary actions to prevent unauthorized access, use, or disclosure.

This includes implementing strong security precautions such as encryption and access limitations, constantly testing systems for security flaws, and periodically educating staff members on data protection best practices. Inadequate PII protection can have major consequences, including damage to one's reputation, monetary losses, and legal liabilities. It is critical for all individuals and organizations to prioritize the protection of PII in order to ensure the security and privacy of people and their information.

6. References: -

[1]. Al-Roubaiey, A., Al-Sadi, J., & Abubaker, M. (2022). A Review of Personal Identifiable Information (PII) and Its Challenges. International Journal of Computer Science and Network Security, 22(6), 32-37.

[2]. European Union Agency for Cybersecurity. (2021). Good practices for data protection. Retrieved from https://www.enisa.europa.eu/publications/good-practices-for-data-protection/good-practices-for-data-protection/view

[3]. National Institute of Standards and Technology. (2021). NIST Privacy Framework: A Tool for Improving Privacy through Enterprise Risk Management. NIST Special Publication 800-37 Rev. 2.

[4]. Office of the Privacy Commissioner of Canada. (2020). What is personal information? Retrieved from https://www.priv.gc.ca/en/privacy-topics/collecting-personal-information/02_05_d_16/

[5]. Office of the Privacy Commissioner of Canada. (2021). Privacy and Cyber Security: Emphasizing Privacy Protection in Cyber Security Activities. Retrieved from https://www.priv.gc.ca/en/privacy-topics/technology-and-privacy/cyber-security-and-privacy/

[6]. Solove, D. J. (2011). Understanding privacy. Harvard University Press.

[7]. U.S. Department of Commerce. (2021). Privacy Shield Framework. Retrieved from https://www.privacyshield.gov/welcome