IJNRD.ORG ISSN : 2456-4184

INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG

An International Open Access, Peer-reviewed, Refereed Journal

Redacting Personal Identifiable Information Using Machine Learning

[1]Aarush Verma
Department of Computer Science and Engineering
PESITM,Shimoga,India

[2]Bhavik Bardia
Department of Computer Science and Engineering
PESITM,Shimoga,India

[3]Lokesh Talluri
Department of Computer Science and Engineering
PESITM,Shimoga,India

[4]Yashwi J Rai
Department of Computer Science and Engineering
PESITM,Shimoga,India

Raghavendra K.
Assistant Professor,
Department of Computer Science and Engineering
PESITM,Shimoga,India

ABSTRACT

Data privacy is increasingly crucial due to the constant occurrence of data breaches. It is imperative to redact personal, sensitive and confidential information from documents to mitigate risks. There have been many cases from past years of data breaches from various popular companies. When the data contains sensitive information, these leaks pose a serious threat. The traditional approach of manually finding and matching millions of words and then redacting is slow and vulnerable. This paper deals with safeguarding of personal identifiable information (PII) in documents by using machine learning techniques for automated redaction. Our machine learning model will be trained on diverse datasets, enabling it to recognize various forms of PII such as names, addresses and identification numbers, across different document types. Thereafter, the information that is identified as PII is redacted. Additionally, the implementation includes a user-friendly interface, allowing users to customize redaction criteria and maintain control over the redaction process.

Keywords Machine Learning,Natural Language Processing, Deep Learning,Named Entity Recognition,spaCy.

I.INTRODUCTION

Data privacy is a widely discussed issue, particularly in light of the recurrent incidents of data breaches and privacy scandals. A data breach occurs when unauthorized parties infiltrate a system to steal or manipulate sensitive or protected data. On the other hand, a data leak refers to the inadvertent exposure of personal information to the public without the individual's knowledge or consent. Personally identifiable information (PII) is information that is when used alone or with other relevant data can identify an individual. Redacting is the process of editing or preparing a document by masking or removing sensitive information often to protect privacy with legal requirements. This involves hiding specific details such as names, addresses or other Personally Identifiable Information (PII), while maintaining the overall context of the document. Redacting PII ensures responsible information sharing while minimizing the risk of privacy breaches and unauthorized use of personal data. Redacting sensitive information helps prevent data leaks by selectively removing or obscuring confidential details, preserving privacy and minimizing the risk of unauthorized access or disclosure. This paper aims to contribute to data privacy and security by providing a scalable and efficient solution for organizations handling large volumes of sensitive information and sets out to revolutionize PII redaction through the strategic integration of advanced technologies such as machine learning, deep learning, and Natural Language Processing (NLP).These helps in redacting personal information efficiently and also preserving the context of document. Text redaction is a common form of redaction that involves the removal or obscuring of specific text or information within a document. This is typically done to protect sensitive or confidential details from being disclosed. Text redaction can be applied manually or using digital tools depending on nature of document. Image redaction involves the modification or removal of specific visual elements within an image to protect sensitive or confidential information. This process is often used in legal, governmental, or corporate settings where visual content needs to be shared or published while ensuring that certain details are not disclosed. Audio redaction involves the modification or removal of specific audio segments within a recording to protect sensitive or confidential information. This process is commonly used to corporate contexts where audio recordings may contain private information that should not be disclosed.

II.OBJECTIVE

Training a natural language processing (NLP) model to accurately identify and classify personally identifiable information (PII) in different kinds of documents such as in text, audio files and user customisation where user can select the information which should be hidden with the help of user interface. To train an accurate NLP model for PII identification, a labeled dataset is compiled, spanning various document types. Data preprocessing enhances dataset quality. A machine learning model like NER is then trained on this dataset to recognize and classify PII entities. Rigorous evaluation ensures accuracy. Once validated, the model is deployed to analyze new documents, enhancing data privacy and security by accurately identifying and categorizing PII.

III. SYSTEM DESIGN

System designed to perform redaction .The spaCy library is used to process text, specifically looking for named entities like person,email, addresses and others. This is achieved by defining a redaction pattern using spaCy's matcher.The identified entities are redacted in the text and the redacted version of the text is stored.The PyMuPDF library (Fitz) is used to open a PDF document and iterate over its pages. The modified PDF is saved to the specified output path.

IV. METHODOLOGY

The system is designed to perform redaction on both plain text and PDF documents. The spaCy-based code processes text, identifies named entities based on a predefined pattern, and replaces them with a redacted version. The PyMuPDF code opens a PDF, extracts text and redacts specific words on each page.

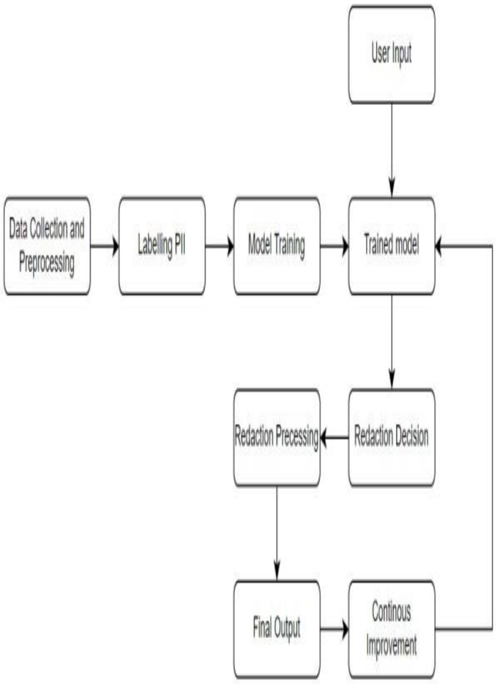


Figure 1 Process of Redaction Using ML

Users provide input to the trained model where the model will be trained for redaction by collecting data, preprocessing and labelling PII. The model is applied for redaction processing resulting in a final redacted output.

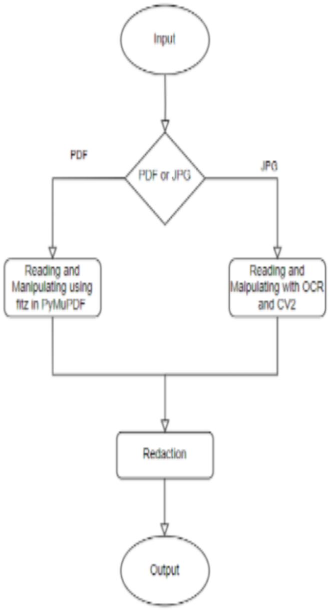
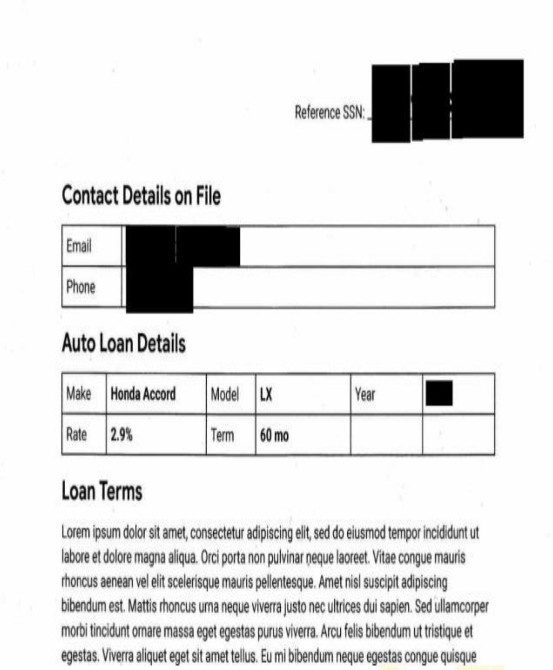

Input
PDF
PDF or JPG
JPG
Reading and Manipulating using fitz in PyMuPDF
Reading and Malpulating with OCR and CV2
Redaction
Output

Figure 2 Flowchart

IV.RELEVANCE IN INDUSTRY

A comprehensive redaction model helps organizations adhere to legal requirements by automatically identifying and redacting sensitive information from various media types. The redaction model can be applied to medical records, images and audio recordings allowing healthcare institutions to share data for research or collaboration while safeguarding patient privacy.A redaction model supports compliance with student privacy laws and facilitates secure sharing of educational resources. Research institutions dealing with multimedia data for studies or experiments can benefit from a redaction model.



Reference SSN: [redacted]

Contact Details on File

Email	[redacted]
Phone	[redacted]

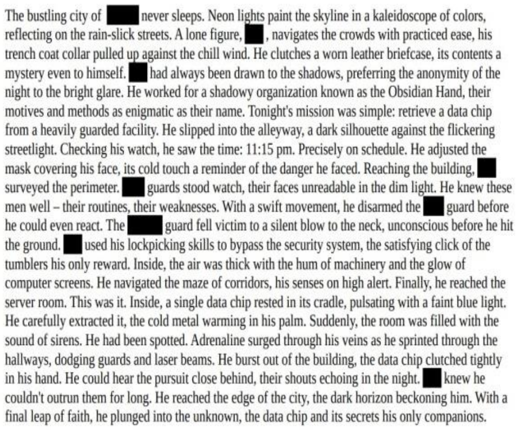
Auto Loan Details

Make	Honda Accord	Model	LX	Year	[redacted]
Rate	2.9%	Term	60 mo		

Loan Terms

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Orci porta non pulvinar neque laoreet. Vitae congue mauris rhoncus aenean vel elit scelerisque mauris pellentesque. Amet nisl suscipit adipiscing bibendum est. Mattis rhoncus urna neque viverra justo nec ultrices dui sapien. Sed ullamcorper morbi tincidunt ornare massa eget egestas purus viverra. Arcu felis bibendum ut tristique et egestas. Viverra aliquet eget sit amet tellus. Eu mi bibendum neque egestas congue quisque

Figure 3 Redacted Document(jpg)



The bustling city of [redacted] never sleeps. Neon lights paint the skyline in a kaleidoscope of colors, reflecting on the rain-slick streets. A lone figure, [redacted], navigates the crowds with practiced ease, his trench coat collar pulled up against the chill wind. He clutches a worn leather briefcase, its contents a mystery even to himself. [redacted] had always been drawn to the shadows, preferring the anonymity of the night to the bright glare. He worked for a shadowy organization known as the Obsidian Hand, their motives and methods as enigmatic as their name. Tonight's mission was simple: retrieve a data chip from a heavily guarded facility. He slipped into the alleyway, a dark silhouette against the flickering streetlight. Checking his watch, he saw the time: 11:15 pm. Precisely on schedule. He adjusted the mask covering his face, its cold touch a reminder of the danger he faced. Reaching the building, [redacted] surveyed the perimeter. [redacted] guards stood watch, their faces unreadable in the dim light. He knew these men well – their routines, their weaknesses. With a swift movement, he disarmed the [redacted] guard before he could even react. The [redacted] guard fell victim to a silent blow to the neck, unconscious before he hit the ground. [redacted] used his lockpicking skills to bypass the security system, the satisfying click of the tumblers his only reward. Inside, the air was thick with the hum of machinery and the glow of computer screens. He navigated the maze of corridors, his senses on high alert. Finally, he reached the server room. This was it. Inside, a single data chip rested in its cradle, pulsating with a faint blue light. He carefully extracted it, the cold metal warming in his palm. Suddenly, the room was filled with the sound of sirens. He had been spotted. Adrenaline surged through his veins as he sprinted through the hallways, dodging guards and laser beams. He burst out of the building, the data chip clutched tightly in his hand. He could hear the pursuit close behind, their shouts echoing in the night. [redacted] knew he couldn't outrun them for long. He reached the edge of the city, the dark horizon beckoning him. With a final leap of faith, he plunged into the unknown, the data chip and its secrets his only companions.

Figure 4 Redacted Document(pdf)

V.CONCLUSION

The project emphasizes the importance of hiding personal information in documents to follow privacy laws before the document is released or shared with others, avoiding legal issues. This Automate Redaction Processes by investing in technologies that automate the redaction process, enhancing efficiency and accuracy while minimizing the potential for human error in handling sensitive information.

REFERENCES

[1]https://venngage.com/blog/data-breach-2022/

[2]https://cloud.google.com/static/dlp/docs/images/redacting-sensitive-data-images-beforeafter.png

[3]https://towardsdatascience.com/revolutionising-redaction-my-final-year-project-fe664e28ef84

[4]https://cloud.google.com/architecture/mlops-continuous-delivery-and-automation-pipelines-in-machine-learning.

[5]https://www.assemblyai.com/blog/what-are-the-top-pii-redaction-apis-and-ai-models/

[6]https://ieeexplore.ieee.org/document/9592788

[7]https://link.springer.com/chapter/10.1007/978-981-19-5689-8_8

[8]https://dl.acm.org/doi/abs/10.1145/3548606.3560572