Proceedings of the Fourth International Conference on Ubiquitous Computing and Intelligent Information Systems (ICUIS-2024)
IEEE Xplore Part Number: CFP24UG3-ART; ISBN: 979-8-3315-2963-5

2024 4th International Conference on Ubiquitous Computing and Intelligent Information Systems (ICUIS) | 979-8-3315-2963-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICUIS64676.2024.10866838

RE-DACT: Adaptive Redaction and Anonymization Tool using Machine-learning

Kishore Kumar I
Student, Department of Information Technology
St. Joseph's College of Engineering, Chennai, India
Kishorekumar21052004@gmail.com

HEPSI AJIBAH A S
Assistant Professor, Department of IT
St. Joseph's College of Engineering, Chennai, India
hepsias@stjosephs.ac.in

***Abstract*—"RE-DACT" is a safe and user-friendly redaction tool for customizable redaction, masking, and anonymization using a user-defined gradational scale. With the help of NLP and ML, the tool gives users the opportunity to specify data elements that are going to be redacted, ranging from simple name removaland more complex anonymization techniques to the generation offully synthetic data, all while keeping the structure of the content unchanged. This tool is usable both online and offline in its web- based interface and takes into account the common input formatsfor text files, images, as well as PDFs. Over time, RE-DACT learns to make realistic synthetic datasets suitable for training, testing, as well as commercial applications without compromising privacy. It therefore holds robust data security with the minimum retention of data and prevents unauthorized access to sensitive information. The secure coding and scalable solutions support real-world applications. Performance is measured in terms ofprecision, recall, F1 score, redaction efficacy, speed, and ease of use.**

"The tool incorporates anonymization techniques, such as k-anonymity and synthetic data generation, to ensure privacy. It makes use of advanced encryption methods and ensures safe placeholders to guard sensitive information from unauthorized access."

***Index Terms*—Redaction, Machine Learning, Anonymization, Data-Security Synthetic Data Multiformat, Natural Language Processing.**

I. INTRODUCTION

Organizations in healthcare, finance, and other legal service sectors have to ensure the safety of confidential information in the wake of increasing sensitivity regarding data and privacy concerns. Redaction involves the removal of specific sensitive information from the documents to ensure compliance with multiple privacy regulations around the globe, including GDPR, HIPAA [13], among others. But with information comingin multiple formats-ranging from CSV files, PDFs to images, manual redaction is often long-winded and error-prone.

This project comes up with an advanced redaction toolcalled RE-DACT that utilizes Natural Language Processing (NLP) algorithms to redact personal and sensitive information across multiple file types-including CSVs, PDFs, and images- using machine learning models. The tool provides customizable levels of redaction from basic to highly advanced. This way, the user can choose the level of protection offered to the data according to their necessity and employs more adaptability and scalability for those industries dealing with sensitive data, enhancing its options for data privacy [9] and security. This flexible and scalable solution can be applied in healthcare, finance, law, or government sectors where reliable document management is important. In addition to the saving in terms of cost of time and resources involved in manual redaction, RE-DACT ensures greater accuracy, which will result in better operational efficiency that the modern challenges in data protection require.

"Tools like Adobe Acrobat Pro and Foxit PDF Editor are well suited for PDFs but not ideal for other formats, such as CSVs and images. Typically, these tools do not provide scalable, secure anonymization techniques, thus leaving sensitive information vulnerable. RE-DACT fills this gap by offering a flexible solution that supports diverse file types and uses advanced machine learning-based redaction and anonymization techniques."

II. LITERATURE SURVEY

Text anonymization has emerged as a critical area of research, driven by the increasing demand for privacy-preserving solutions in data sharing and processing. This review examines recent advancements in text anonymization techniques and the associated risks of re-identification, highlighting contributions from key studies. [1] developed a machine learning-based approach to assess the disclosure risks of anonymized documents. Their study demonstrated the vulnerabilities of anonymized text when subjected to re-identification attacks, emphasizing the need for robust anonymization strategies. Similarly, [2] presented a semi-automated system for redacting sensitive information using machine learning, which facilitated efficient anonymization while maintaining document utility. [3] proposed a deep learning-based framework for evaluating the privacy-utility tradeoff in anonymized documents. This method provided an automatic and scalable evaluation mechanism, underscoring the potential of AI-driven solutions in privacy assessment. [4] further explored AI-driven anonymization, introducing techniques that combined detection and anonymization of sensitive text, which showed promise for real-world applications. [5] provided a comprehensive review of anonymization models for text data, outlining state-of-the-art methods and identifying key challenges, such as balancing privacy with data utility. They also highlighted future research directions, including the integration of contextual understanding in anonymization models. Patsakis and [6] examined the efficacy of human-driven versus machine-driven anonymization methods in the context of large language models, revealing that machines often outperformed humans in text anonymization tasks while maintaining scalability and efficiency.

979-8-3315-2963-5/24/$31.00 ©2024 IEEE

[7] addressed the challenges of anonymizing electronic health record (EHR) data, proposing strategies that combined de-identification techniques with anonymization frameworks. Their study emphasized the importance of domain-specific approaches for preserving patient privacy in healthcare research.

III PROPOSED METHODOLOGY

Two key steps involve the creation of the ML-based redaction tool: NER and implementation of machine learning models to identify and redact sensitive information. In the case of CSV files, the tool uses the pandas library to extract data from the columns, identifies continuous text, names, addresses, phone numbers, etc. The spaCy ML tool applies NER for detecting personal or sensitive information. A key role is played in this process by the NLP (Natural Language Processing), which enables the tool to understand and to interpret the extracted text's context. It allows for finding different forms of sensitive data, although expressed in a different way or ambiguously, like dates, social security numbers, and other personal identifiers. Once the information is identified, the tool replaces sensitive information with generic terms, dummy strings, or obscures the information to ensure privacy, depending on the user's chosen level of redaction.

Extracting text for images begins with using Tesseract OCR on embedded images. Then, the text is processed by NER models, where checks for any sensitive content are made. The tool will then obscure, mosaic, or erase the detected regions of the photograph to ensure that sensitive information is fully protected. For PDF files, extracted text is done using tools such as PyPDF2 or pdf-plumber. The extracted text is then treated similarly to CSV files and processed with NER to identify and redact any sensitive data, thereby ensuring complete redaction across all supported file types.

A. Data Collection and Pre-Processing

For effective data processing in the RE-DACT tool, several steps are followed to upload different types of files such as CSV, PDF, and images into a state that is ready for redaction and anonymization. The process starts with file parsing and text extraction. In the case of CSV files, the system extracts headers and key data types from each column, identifying fields that may contain sensitive information, such as names, addresses, or phone numbers. In the case of PDFs and images, Optical Character Recognition (OCR) in Fig. 2 is used to extract text from an embedded image or scanned document.

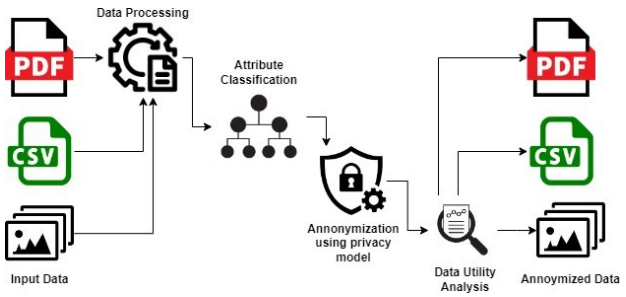

Fig 1. Architecture diagram

Next, it deploys NER models to extract features from the sensitive entities of personal identifiable information (PII) and even financial data or medical records. Data cleaning follows to remove unwanted characters, standardize formats, and make the dataset consistent. Missing values are flagged and dealt with in accordance with the privacy choices by the user, either by replacement or redaction . By including data from open-source repositories like Kaggle, offering fully representative and diverse datasets, the tool can handle a wide variety of data very efficiently. The pipeline for data processing detects sensitive information correctly in different file formats, thus increasing precision within the redaction process. This structured approach makes sure that the redaction and anonymization of data happen consistently and efficiently through RE-DACT, while protecting the sensitive information.

B. IDENTIFICATION

Natural Language Processing (NLP): NLP models are applied over the extracted text to look for patterns that may correspond to the categories of sensitive data: personal identifiers (including, but not limited to, names, phone numbers, and email addresses); financial information (including, but notlimited to, bank account and credit card numbers); and medical records including diagnoses and other conditions.

NER is an important component of NLP, and it is used to identify sensitive entities in the text and assign them category tags. This task makes use of pre-trained models such as spaCy, NLTK.

Custom Regex Rules: There are custom regular expressions for data types, including dates, Social Security Numbers, and telephone numbers, which have been built to improve the identification process, through different formats.

C. ALGORITHM

A redaction tool, with its backbone in machine learning, uses multiple algorithms to improve the detection and redaction of sensitive information across file formats. It uses named entity recognition to identify the sensitive information with NLP-based techniques where NLTK and Tesseract OCR are used. PyPDF2 and pdf-plumber tools pick up text from the PDFs for anonymization, and regular expressions help in targeted redaction. Randomized string replacement masks sensitive information while simple methods like Gaussian blur will obscure the content in an image.

Techniques foranomaly detection, such as Isolation Forest and Support Vector Machines, assist in the discovery of unknown sensitive data, thus effecting proper maintenance of privacy. Lastly, there areadvanced features like encrypted redaction. This applies secureand reversible encryption techniques that will only decrypt the sensitive information under authorized conditions, thus increasing security on all file types of the managed tool.

979-8-3315-2963-5/24/$31.00 ©2024 IEEE

D. MACHINE LEARNING MODELS

A redaction tool employs develop machine ML models that can effectively identify and later delete sensitive information. Tesseract is used for Optical Character Recognition so that there will be extraction of text from images, whereas spaCy provides named entity recognition, enabling identification of sensitive entities within extracted text. For tabular data, Pandas will ease manipulation and traditional machine learning models such as Random Forests or SVMs could be applied to classify sensitive entries in CSV files. NLP models, such as spacy can be implemented for image processing tasks, and this enhances the accuracy of redaction on various data formats

E. MODEL EVALUATION

Major metrics in this case include precision-the rate at which the correct identification of sensitive information takes place-and recall, which shows how well the system detects all the sensitive information. Other major performance metrics include false positives and false negatives, the former taking place when nonsensitive data is misidentified as sensitive while the latter when a system fails to recognize sensitive information. Redaction accuracy is the overall correctness measure, while latency denotes the speed of redaction, which can be critical for real-time applications. So, these two sets of metrics together ensure that the model should redact sensitive information properly as well as it must be correct and efficient.

IV. RESULTS AND DISCUSSIONS

Finally, the performance metrics for the redaction tool are measured on accuracy, precision, and recall. Accuracy refers to the percentage of sensitive items that were correctly identified. Precision is equated to true positives or, rather, the redactions of all identified sensitive entries. This reflects the proportion of redacted items identified against the actual number of sensitive items. Besides this, image and PDF redaction quality is visually checked for secure sufficient coverage of confidential information, and textual evaluation is done on verifying whether the sensitive entries are correctly masked for CSV files. Usability and effectiveness assessment are done with userfeedback and error analysis. If the manual process does not satisfy the expectations of the user, then remediation occurs. Other performance measures that include processing time and the model's robustness are also shown to confirm that thetool functions effectively and generalizes well across different datasets. Overall, such evaluations enhance the effectiveness of the tool in offering robust information sensitivity protection.

979-8-3315-2963-5/24/$31.00 ©2024 IEEE

document	text	tokens	trailing_wh	labels	prompt	prompt_id	name	email	phone	job	address	username	url	hobby
1073d46f-	My name i	['My', 'nan	[True, True	['O', 'O', 'O		1	Aaliyah Po	aaliyah.po	(95) 94215	jeweler	97 Lincoln Street			Podcastin
5ec717a9-	My name i	['My', 'nan	[True, True	['O', 'O', 'O		1	Konstantin	konstantin	0475 4429	developer	826 Webster Street			Quilting
353da41e-	As Mieko M	['As', 'Miek	[True, True	['O', 'B-NA		3	Mieko Mit	mieko_mit	+27 61 222	account m	1309 Southwest 71st Terrace			Metal det
9324ee01-	My name i	['My', 'nan	[True, True	['O', 'O', 'O		1	Kazuo Sun	kazuosun@	0304 2215	air traffic c	736 Sicard Street Southeast			Amateur r
971fe266-	My name i	['My', 'nan	[True, True	['O', 'O', 'O		3	Arina Sun	arina-sun@	0412 1245	dental hyg	5701 North 67th Avenue			Related
13166f5c-	Baha Hoff	['Baha', 'H	[True, True	['B-NAME_		2	Baha Hoff	bahahoffm	+27 68 67(	lawyer	45 Baldridge Road			Related
88db5573-	From the r	['From', 'th	[True, True	['O', 'O', 'O		0	Natalia Gr	nataliagro:	(98) 96894	waitress	6420 Via Baron			Related
8c9c705b-	My name i	['My', 'nan	[True, True	['O', 'O', 'O		1	Alexander	alexandert	+86 10746	saleswom	1890 Orchard View Road			Metalwor
68b043f5-	My name i	['My', 'nan	[True, True	['O', 'O', 'O		1	Kuo Lopez	kuolopez@	+27 49 207	professor	4188 Summerview Drive			Kite surfir
a9c3689c-	Hi, I'm Ash	['Hi,', "I'm"	[True, True	['O', 'O', 'B		7	Ashok Ma	ashokma5	0932 173 5	developer	3763 Laure	@ashok.m	https://wv	Jigsaw pu

Fig. 2: Sample CSV data set

a. Performance

The efficiency of the proposed redaction tool was evaluated across three file types: CSV, PDF, and image files, with a focus on performance differences compared to existing models. These results reveal distinct strengths and areas for improvement, highlighting the tool's adaptability and efficiency across diverse file formats (Table 1).

CSV Files

The redaction tool demonstrated exceptional performance in processing CSV files, handling structured tabular data with both speed and accuracy. For 100KB files, the tool completed redaction tasks in just 2 seconds, scaling linearly to 14 seconds for 10MB files (Fig. 3). This reflects the inherent simplicity and efficiency of redacting structured data. In contrast, most existing tools lack optimization for structured data, resulting in slower processing and reduced accuracy.

PDF Files

For PDF files, the tool showed strong performance with simple, text-based PDFs, achieving redaction times as low as 3 seconds for 100KB files and up to 20 seconds for 10MB files (Fig. 4). However, when processing complex PDFs containing multiple columns, images, or scanned text requiring OCR, performance was impacted. Despite this, the tool maintained a notable advantage over existing models, which often require up to 50 seconds for redacting similar 10MB files. This demonstrates the tool's capability to adapt to file complexity while maintaining competitive efficiency.

Image Files

In the case of image files, where OCR is crucial, the tool processed 100KB images in approximately 5 seconds and scaled to 35 seconds for 10MB images (Fig. 5). Performance degradation was linear, showcasing predictable scaling. By contrast, existing models exhibited significant performance drops with larger or higher-resolution images, with processing times for 10MB files reaching up to 50 seconds. This highlights the tool's superior ability to maintain consistent performance across file sizes.

File Size (KB)	CSV Processing Time (Your Tool)	PDF Processing Time (Your Tool)	Image Processing Time (Your Tool)	Existing Tool (PDF/Image) Processing Time
100	2	3	5	10
1000	4	6	10	15
2000	6	9	15	20
4000	8	12	20	30
6000	10	15	25	35
8000	12	18	30	40
10000	14	20	35	50

Table 1: Performance table for proposed model vs traditional model

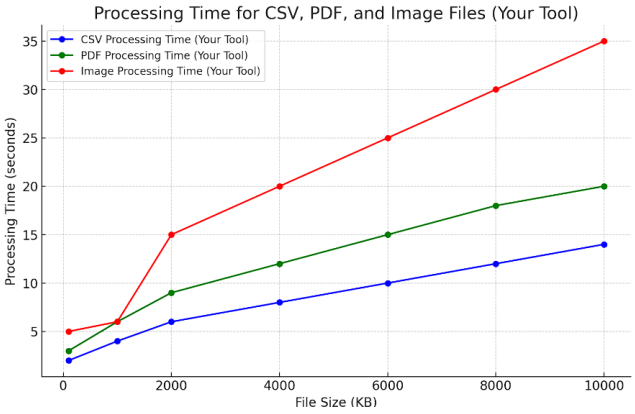

Fig 3: Performance based Processing Time

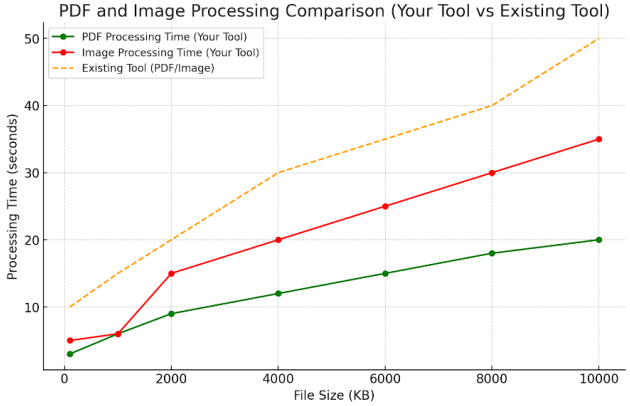

Fig 4: Existing and Proposed tool performance Analysis

979-8-3315-2963-5/24/$31.00 ©2024 IEEE

b Over-Redaction of PDF, IMAGES, CSV

This redaction tool goes further than the basic data formats, for example, to support PDF and images really robustly. It makes sure that comprehensive coverage is given for several types of information. In the case of PDF, this tool has efficiently identified all confidential information such as names, addresses, and other personal identifiers so that it could keep the integrity of the document intact but nevertheless uphold compliance with all regulations on privacy. That is, the use of sophisticated algorithms for detecting and blackout any visually sensitive information, such as faces or license plates, enables a user to protect himself to the utmost extent. Thus, this type of multiple format capability, while making the tool versatile in its applicability, also enhances the ability of users to control and safeguard information in its most sensitive

c Deployment of the model

The redaction tool is a user-friendly application allowing users to upload data easily, which could be in the form of PDFs, CSV (Fig 7), images, and others. Users may also specify the attributes to be redacted, such as personal identifiers or confidential information. All these can be processed quicklyby the system to mask or remove selected contents for that specific file, making sure the sensitive information is taken care of after the redaction. Users can then download the redacted files in their original format. It was designed to be efficient with large data amounts and intuitive forskill levels. In addition, the tool has online and offline functionalities to provide a reliable solution that fits the needs of individuals and organizations when properly protecting sensitive data. forms, making it an invaluable resource for those handling confidential data (Fig 5 and Fig 6) for output.

Fig 5 Output image 1

Fig 6 Output image 2

rompt_id	name	email	phone	job	address	username
1	[REDACTE	[REDACTE	[REDACTE	jeweler	[REDACTE	nan
1	[REDACTE	[REDACTE	[REDACTE	developer	[REDACTE	nan
3	[REDACTE	[REDACTE	[REDACTE	account m	[REDACTE	nan
1	[REDACTE	[REDACTE	[REDACTE	air traffic c	[REDACTE	nan
3	[REDACTE	[REDACTE	[REDACTE	dental hyg	[REDACTE	nan
2	[REDACTE	[REDACTE	[REDACTE	lawyer	[REDACTE	nan
0	[REDACTE	[REDACTE	[REDACTE	waitress	[REDACTE	nan
1	[REDACTE	[REDACTE	[REDACTE	saleswom	[REDACTE	nan

Fig 7 Sample Anonymized CSV Output

V CONCLUSION

In conclusion, the redaction tool is an extremely important resource to manage and protect sensitive information in all forms-from CSVs to PDFs, even images. That makes the task of finding, deleting, and masking confidential data easier with a friendly interface as well as excellent functionalities in powers. If adaptable to various kinds of data, then the people or organizations in control of such information can be com- pliant with privacy regulations yet maintain their information safe. With this increasing demand for privacy protection, this redaction tool becomes a much-needed solution for any data seeker to try and secure his own against unauthorized access.

979-8-3315-2963-5/24/$31.00 ©2024 IEEE

REFERENCES

[1] Manzanares-Salor, Benet, David Sánchez, and Pierre Lison. "Evaluating the disclosure risk of anonymized documents via a machine learning-based re-identification attack." Data Mining and Knowledge Discovery 38, no. 6 (2024): 4040-4075.

[2] Cumby, Chad, and Rayid Ghani. "A machine learning based system for semi-automatically redacting documents." In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 25, no. 2, pp. 1628-1635. 2011.

[3] Manzanares Salor, Benet. "Automatic privacy and utility evaluation of anonymized documents via deep learning." Master's thesis, Universitat Politècnica de Catalunya, 2023.

[4] Böhlin, Felix. "Detection & Anonymization of Sensitive Information in Text: AI-Driven Solution for Anonymization." (2024).

[5] Lison, Pierre, Ildikó Pilán, David Sánchez, Montserrat Batet, and Lilja Øvrelid. "Anonymisation models for text data: State of the art, challenges and future directions." In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 4188-4203. 2021.

[6] Patsakis, Constantinos, and Nikolaos Lykousas. "Man vs the machine in the struggle for effective text anonymisation in the age of large language models." Scientific Reports 13, no. 1 (2023): 16026.

[7] Kushida, Clete A., Deborah A. Nichols, Rik Jadrnicek, Ric Miller, James K. Walsh, and Kara Griffin. "Strategies for de-identification and anonymization of electronic health record data for use in multicenter research studies." Medical care 50 (2012): S82-S101.

979-8-3315-2963-5/24/$31.00 ©2024 IEEE