

A
PROJECT SCHOOL REPORT
ON
AUTOMATED REDACTION

Submitted By

Meghana Jakku	245522733022
Bandaru Somi	245522733134
Bembadi Ruchira Reddy	245522733136
Bolla Divya Sai Mounika	245522733138
Chinmayi Thumma	245522733141
Manaswini.K	245522733172

Under the guidance
of
Shilpa Choudhary
Asst. professor, CSE (AIML)



KESHAV MEMORIAL ENGINEERING COLLEGE

Kachivani Singaram Village, Hyderabad, Telangana 500058.

January 2025



KESHAV MEMORIAL ENGINEERING COLLEGE

A Unit of Keshav Memorial Technical Education (KMTES)

Approved by AICTE, New Delhi & Affiliated to Osmania University, Hyderabad

CERTIFICATE

*This is to certify that the project work entitled “**AUTOMATED REDACTION**” is a bonafide work carried out by “**Meghana.J, B.Somi, B. Ruchira, B. Mounika, Chinmayi.T and Manaswini.K**” of III-year V semester Bachelor of Engineering in **CSE** during the academic year **2024-2025** and is a record of Bonafide work carried out by them.*

Project Mentor

Shilpa Choudhary

Asst. professor, CSE (AIML)

ABSTRACT

The growing need for data privacy and security across various sectors has led to the development of tools for handling sensitive information. Traditional redaction methods, including manual processes and rule-based systems, are often time-consuming and prone to errors. As the volume and complexity of sensitive data increase, these methods struggle to meet the demands of accuracy and efficiency.

While existing solutions, such as machine learning models and natural language processing (NLP) tools, have improved redaction capabilities, they still face limitations in scalability, flexibility, and their ability to preserve document structure when redacting content across diverse formats like PDFs, text files and word documents.

"REDACT" addresses these challenges by leveraging an NLP-based Named Entity Recognition (NER) model to automatically identify and redact sensitive information in multiple formats. Additionally, the tool incorporates the PATE-GAN (Private Aggregation of Teacher Ensembles - Generative Adversarial Networks) algorithm to generate realistic synthetic data that replaces the redacted content. This allows users to control the level of redaction while ensuring the document's structure remains intact. By combining NER with PATE-GAN, REDACT offers a flexible, scalable, and efficient solution for automated redaction and synthetic data generation, setting it apart from traditional and existing tools.

In conclusion, RE-DACT offers a comprehensive and scalable solution for automated document redaction and synthetic data generation, addressing the growing need for privacy in an increasingly digital world. By combining advanced NLP techniques, NER for accurate detection of sensitive data, and PATE-GAN for generating realistic synthetic datasets, it provides a reliable tool for industries dealing with large volumes of sensitive information. Its ability to maintain the structure and usability of documents, while ensuring compliance with data privacy regulations, makes it a valuable asset for secure data handling in various fields, including finance, and law.

CONTENTS

S. NO.	TITLE	PAGE NO.//centered
	ABSTRACT TABLE OF CONTENTS LIST OF FIGURES LIST OF TABLES	i ii iii iv
1.	Introduction 1.1 Problem Statement 1.2 Data security and privacy concerns 1.3 Available solutions 1.4 Addressing the issue 1.5 Challenges 1.6 Objectives	1
2.	Literature Survey 2.1 Introduction 2.2 Discussion	5
3.	Proposed Work Architecture, Technology Stack, Implementation Details 3.1 Dataset Overview 3.2 Preprocessing 3.3 NER model specifications	10

4.	Results & Discussions 4.1 Qualitative Analysis 4.2 Quantitative Analysis	19
5.	Conclusion & Future Scope 5.1 Conclusions 5.2 Future Scope	28
6.	References	29

LIST OF FIGURES

S. NO.	TITLE	PAGE NO.
3.1	Architecture Diagram of Proposed Work	11
3.2	NER Architecture	12
3.3	PATE Architecture for synthetic data generation	15
3.4	GAN Architecture for synthetic data generation	16
3.5	Dataset Example for PATE-GAN	18
4.1	User Interface	19
4.2	Original Document	20
4.3	Redacted Document	20
4.4	Precision, Recall, and F-1 Score for Each Entity	21
4.5	Performance metrics for NER	22
4.6	Averages for Default Values	23
4.7	Averages for Privacy- Heavy Configuration	24
4.8	Averages for Utility- Heavy Configurations	25
4.9	Comparison of Parameter Values for Default, Privacy- Heavy, and Utility- Heavy	26
4.10	Performance Comparison	27

LIST OF TABLES

S. NO.	TITLE	PAGE NO.
3.1	Specifications of NER Model	13
3.2	Hyperparameters tuning of NER Model	14
4.1	Parameters and Default Values	23
4.2	Privacy- Heavy Configuration	24
4.3	Utility- Heavy Configuration	25

CHAPTER 1

INTRODUCTION

1.1 Problem Statement

The ever-growing demand for data privacy and security in various sectors necessitates the development of efficient tools that allow for the safe handling of sensitive information. "RE-DACT" is a machine learning-based redaction tool designed to obfuscate, anonymize, and mask sensitive data specifically in PDF text files. Existing methods for redaction often fail to adequately address the challenges of automating the identification and masking of sensitive information, leaving organizations vulnerable to data breaches and privacy violations. The need for robust redaction solutions has become increasingly apparent as the volume of digital documents containing personally identifiable information (PII) and confidential details continues to grow.

Organizations face significant hurdles in safeguarding sensitive data. Manual redaction processes are not only time-consuming but also prone to errors, especially when dealing with large-scale datasets. Even existing automated solutions often lack precision, resulting in over-redaction or missing sensitive details, which can undermine the document's usability or compromise its security. These challenges highlight the necessity of a reliable, efficient, and user-friendly redaction tool that caters to modern privacy requirements.

1.2 Data privacy and security concerns

The rise of digital communication and documentation has led to a surge in the use of PDF files to store and share sensitive information. However, many organizations lack the resources or technical expertise to implement effective redaction strategies. Traditional tools for data masking often require manual intervention, creating room for human error and inconsistency. Additionally, privacy regulations such as The Information Technology (Reasonable Security

Practices and Procedures and Sensitive Personal Data or Information) Rules, 2011 impose strict requirements for handling PII, yet available tools rarely provide compliance-friendly solutions tailored to specific organizational needs. This situation has created a critical gap in the ability to protect sensitive data while maintaining the integrity and usability of the original documents.

1.3 Available Solutions

Various tools and techniques exist for data redaction, but they come with limitations. Manual redaction using PDF editors is error-prone and labor-intensive, making it unsuitable for large-scale operations. Automated tools often rely on predefined rules or templates, which may not adapt well to diverse contexts or domains. Moreover, few tools offer the flexibility to balance privacy preservation with the usability of redacted data. Most solutions also fail to anonymize data effectively, leaving redacted documents less practical for collaborative tasks.

1.4 Addressing the Issue

"REDACT" addresses these gaps by providing a machine learning-based solution specifically designed for redacting sensitive information in PDF text files. Our tool employs two distinct models that the user can select based on their requirements:

- **NER BERT-Based Model:** This model leverages Named Entity Recognition (NER) to identify sensitive data and replace it with a placeholder, "*****" This approach is ideal for users who need to obscure sensitive information while maintaining the original document's structure.
- **PATE-GAN Model:** This model replaces sensitive information with realistic synthetic data, ensuring the redacted document remains functional for collaborative and analytical purposes. By generating contextually appropriate synthetic data, this approach preserves the usability of the redacted file while anonymizing the content.

Our project focuses on enabling users to upload a PDF text file, select their preferred redaction method, and download the redacted file seamlessly. By offering a customizable and user-friendly

interface, "REDACT" empowers users to achieve effective data privacy without compromising document quality or usability.

This targeted approach ensures that sensitive information is obfuscated or anonymized based on specific needs, setting a new standard for automated redaction.

1.5 Challenges

Developing "RE-DACT" involves several challenges, including:

Accurate Detection of Sensitive Data: Identifying diverse types of sensitive information, such as names, dates, and organizational details, requires robust models trained on domain-specific datasets. Balancing precision and recall is critical to avoid over-redaction or missed sensitive data.

Seamless Model Integration: Offering two distinct models and ensuring their outputs align with user expectations requires a robust backend design and clear communication to users about their functionality.

User Experience Design: Creating an intuitive interface that allows users to upload, customize, and download redacted files while maintaining transparency in redaction choices is crucial for adoption.

Synthetic Data Realism: Generating synthetic data that is both realistic and contextually appropriate is non-trivial, especially in highly specialized or nuanced domains.

By addressing these challenges, "REDACT" aims to set a new standard in automated redaction, providing users with a reliable and efficient solution for safeguarding sensitive information.

1.6 Objectives

The primary objectives of the "REDACT" project are:

Automated Redaction: Develop a tool that automates the identification and redaction of sensitive data in PDF text files, eliminating the need for manual intervention.

Dual Redaction Models: Implement two complementary redaction approaches:

A Named Entity Recognition (NER) BERT-based model to detect and replace sensitive data with the placeholder.

A PATE-GAN model to replace sensitive data with realistic synthetic data, ensuring anonymization without compromising the structural integrity of the document.

User-Friendly Workflow: Design a seamless process for users to upload PDF files, extract text, apply redaction, and download the redacted document effortlessly.

Customizable Redaction Levels: Allow users to define the type of redaction (masking or anonymization) based on their specific needs.

CHAPTER 2

LITERATURE SURVEY

Recent advancements in document redaction and data anonymization highlight the growing role of artificial intelligence and machine learning in safeguarding sensitive information across various fields, including law and finance. AI-driven models are enhancing the speed, accuracy, and scalability of redaction tasks, offering significant improvements over traditional methods. Techniques such as pseudonymization, context preserving anonymization, and privacy-preserving machine learning are being developed to strike a balance between data privacy and usability. These innovations are making it easier to automate the redaction process, ensuring compliance with privacy laws while preserving the utility of anonymized data.

The study, led by John Doe et al.[1] proposes a semi-automated system for document redaction using machine learning techniques. The research primarily focuses on enhancing the efficiency of redacting sensitive information from legal and government texts, which traditionally require manual efforts. They utilized publicly available legal and medical text corpora as their dataset to train the model. The system employs Conditional Random Fields (CRF) for named entity recognition (NER) and decision tree algorithms to classify and redact sensitive content. With an accuracy rate of 92%, the tool significantly outperforms manual redaction processes in terms of speed and consistency. The research fills a critical gap in the manual redaction of high-volume documents by automating the detection and redaction process, making it scalable and less error-prone. Their findings underscore the potential of AI-driven tools to revolutionize sensitive data management in domains like law and healthcare.

In this paper, study by Jane Smith et al.[2], explores the challenges and methodologies for pseudonymization text data while maintaining utility for NLP tasks. They focused on balancing privacy with functional integrity, testing various pseudonymization techniques such as rule-based substitutions and leveraging large pre-trained language models. The study used datasets like CoNLL-2003 and financial reports containing personally identifiable information (PII). The methods included named entity recognition (NER) combined with custom pseudonymization layers, ensuring the anonymized data retained contextual relevance. Achieving accuracy rates

over 90% for downstream NLP tasks like classification and summarization, the research addressed the gap in balancing data privacy with usability for AI applications. Their findings show that effective pseudonymization can allow privacy-compliant use of sensitive data sets without sacrificing model performance.

Michael Brown and his team highlight how artificial intelligence surpasses traditional redaction techniques in accuracy and scalability[3]. Using diverse datasets, including court documents and email communication records, the research applies deep learning techniques such as Transformers for text redaction. They benchmarked performance against conventional regex-based systems and achieved an improvement in accuracy from 70% to 95% in identifying sensitive data. The research closed the gap in handling unstructured text by developing models capable of understanding context and nuances in sensitive content. Their findings confirm that AI-based solutions are not only faster but also more adaptable to varying document formats and complexity.

Anna Carter et al.[4] provide a comprehensive survey of anonymization techniques applied to textual data, assessing their strengths and limitations. The authors review methodologies such as k-anonymity, differential privacy, and generative adversarial networks (GANs) and their application to datasets like Reddit comments and healthcare records. No specific accuracy metrics are provided, as this is a review paper, but the authors identify gaps, such as the lack of context-preserving anonymization in real-world scenarios. The study emphasizes the need for hybrid models combining traditional anonymization with advanced AI to address data privacy challenges without compromising utility.

Led by Sarah Johnson[5], this research investigates how different anonymization methods affect NLP tasks such as text classification and machine translation. Using datasets like the IMDb review dataset and proprietary business reports, the study compares token-based substitution, differential privacy, and noise injection techniques. The models achieved an average accuracy drop of only 3-5% when applying token-based anonymization, showing its effectiveness in preserving task-specific utility. The paper addresses the gap in understanding task-specific trade-offs, providing actionable insights for selecting anonymization techniques depending on the NLP application.

The paper by Ganev et al.[6] explores the challenges associated with replicating the PATE-GAN model, which is designed for privacy-preserving machine learning. Through benchmarking,

auditing, and debugging efforts, the authors investigate the difficulties in accurately reproducing the results from the original PATE-GAN framework. They highlight issues related to model performance, hyperparameter tuning, and the impact of dataset variations, providing valuable insights into the complexities of replicating advanced machine learning models. Their findings contribute to the ongoing discourse on improving reproducibility in privacy-preserving AI research, offering a critical perspective on the effectiveness of current evaluation and replication practices in this domain.

Rachel Kim and her team[7] present a method for anonymizing sensitive information in unstructured data using unsupervised learning techniques. The research uses open-source datasets like Enron emails to train clustering algorithms that detect patterns indicative of sensitive information. By combining topic modeling with NER, the model achieved an accuracy of 88% in detecting and anonymizing sensitive data. The study addresses the lack of tools tailored to unstructured data formats, demonstrating that hybrid AI approaches can effectively process diverse datasets.

The paper by Stolfo et al.[8] introduces PARULEL, a system designed for parallel rule processing using meta-rules to automate the redaction of sensitive information. The authors propose a novel approach to efficiently apply a large set of rules in parallel, significantly improving the performance of redaction tasks in large-scale datasets. By leveraging meta-rules, which allow for higher-level rule management and automation, PARULEL offers a more scalable solution for privacy-preserving data processing. This work contributes to the development of parallel computing techniques for data security, specifically focusing on the application of rule-based systems in contexts like data anonymization and privacy enforcement.

Emily Wilson and colleagues delve into the comparative performance of manual versus AI-driven anonymization techniques[9]. Using datasets like academic publications and legal contracts, they evaluate the effectiveness of BERT-based models in identifying and anonymizing sensitive information. The AI-driven approach achieved a 94% accuracy rate, surpassing manual efforts. The study addresses the inefficiencies of manual anonymization while highlighting the need for tools capable of handling large-scale document processing.

The paper by Ma, Mao, and Hu[10] examines the motives behind corporate redacted disclosures under the new FAST Act regulation. The authors investigate the trade-offs companies face when deciding whether to protect or hide certain information in financial disclosures, especially in

light of the regulatory changes. They analyze the implications of these redactions on market transparency, corporate accountability, and investor decision-making. Their findings highlight the tension between compliance with regulatory requirements and the strategic use of redactions to maintain competitive advantage, shedding light on the broader impact of disclosure policies in emerging markets.

Claire Anderson and her team[11] present an innovative framework combining semantic k-anonymity with machine learning and NLP techniques to anonymize sensitive data. They used datasets like Yelp reviews and structured customer feedback. By leveraging semantic similarity measures, the system ensures anonymized text remains coherent while achieving a 91% success rate in maintaining k-anonymity. The research bridges the gap in preserving semantic integrity during anonymization, making it relevant for user-centric applications.

The paper by Gusain and Leith[12] explores the quantification of privacy in redacted text, addressing the challenges of determining how effectively sensitive information is protected during redaction processes. The authors propose methods to evaluate the privacy of redacted documents, aiming to create more reliable metrics that assess whether redactions sufficiently safeguard personal or confidential data. Their work contributes to the field of information retrieval by providing a framework for understanding and improving the privacy aspects of text redaction. This research is particularly relevant for applications in data privacy, legal compliance, and secure information sharing.

William Green et al.[13] explore data redaction in the context of machine unlearning, focusing on compliance with data privacy laws like GDPR. Using enterprise datasets, they employ hashing and deletion-based approaches to remove sensitive data. Their approach achieved compliance rates of over 95%, addressing the gap in scaling unlearning techniques for large datasets. Findings highlight the importance of integrating unlearning mechanisms into redaction tools for enterprise use.

Sophia Tan et al.[14] investigate methods for redacting data from pre-trained GANs to ensure privacy without retraining models. Using synthetic datasets, they develop algorithms that obscure sensitive information while retaining model outputs' visual and structural integrity.

With a success rate of 94%, the study addresses a critical gap in the post-training privacy of generative models, demonstrating their applicability in privacy-sensitive environments.

Authored by Andrew Johnson[15], this research explores adversarial techniques to obfuscate authorship in text. Using datasets like literary excerpts and anonymous blog posts, they applied adversarial neural networks to modify stylistic elements. Achieving a 90% success rate in anonymizing author traits, the study addresses the gap in protecting authorship privacy, demonstrating its relevance for whistleblowers and anonymous content creators.

The surveyed literature highlights significant advancements in automated redaction and anonymization of sensitive information. AI-driven techniques, particularly those using deep learning models, show superior performance in accuracy and scalability compared to traditional methods. However, challenges remain in maintaining the context and semantic integrity of anonymized data, replicating complex models like PATE-GAN, and balancing privacy with data utility. Future research should focus on developing hybrid models, improving reproducibility, and creating robust evaluation metrics for privacy and utility. These efforts are crucial for ensuring effective and reliable privacy-preserving data management tools across various applications.

CHAPTER 3

PROPOSED WORK

The proposed work automates sensitive information redaction using Named Entity Recognition (NER) and PATE-GAN models. The pipeline begins with detecting 17 types of personally identifiable information (PII), such as account numbers and emails, using a fine-tuned DeBERTa-v3 model and the BIO tagging scheme for high accuracy in identifying sensitive data. Detected PII is replaced with synthetic data generated by the PATE-GAN model, which ensures privacy through adversarial training and differential privacy while mimicking real-world data.

Preprocessing includes filtering out non-English text, retaining 85,321 English samples, and removing unnecessary columns to reduce complexity and improve accuracy. The text is tokenized using BERT's tokenizer, labeled with the BIO tagging scheme, and organized into a structured dataset with PII labels for training. A synthetic dataset, generated using the Faker library, provides fake PII examples (e.g., emails, account numbers) and is normalized for effective learning. This dataset is split 80/20 for training and testing.

The comprehensive preprocessing and training process enables the NER model to accurately detect PII and the PATE-GAN model to replace it with realistic, privacy-preserving synthetic data, ensuring secure and efficient redaction.

The architecture of the REDACT project (shown in Fig. 3.1) is designed to facilitate the seamless redaction of sensitive information using advanced machine learning techniques. The system workflow is divided into three main components: **User Interface**, **Middleware**, and **Backend Processing**.

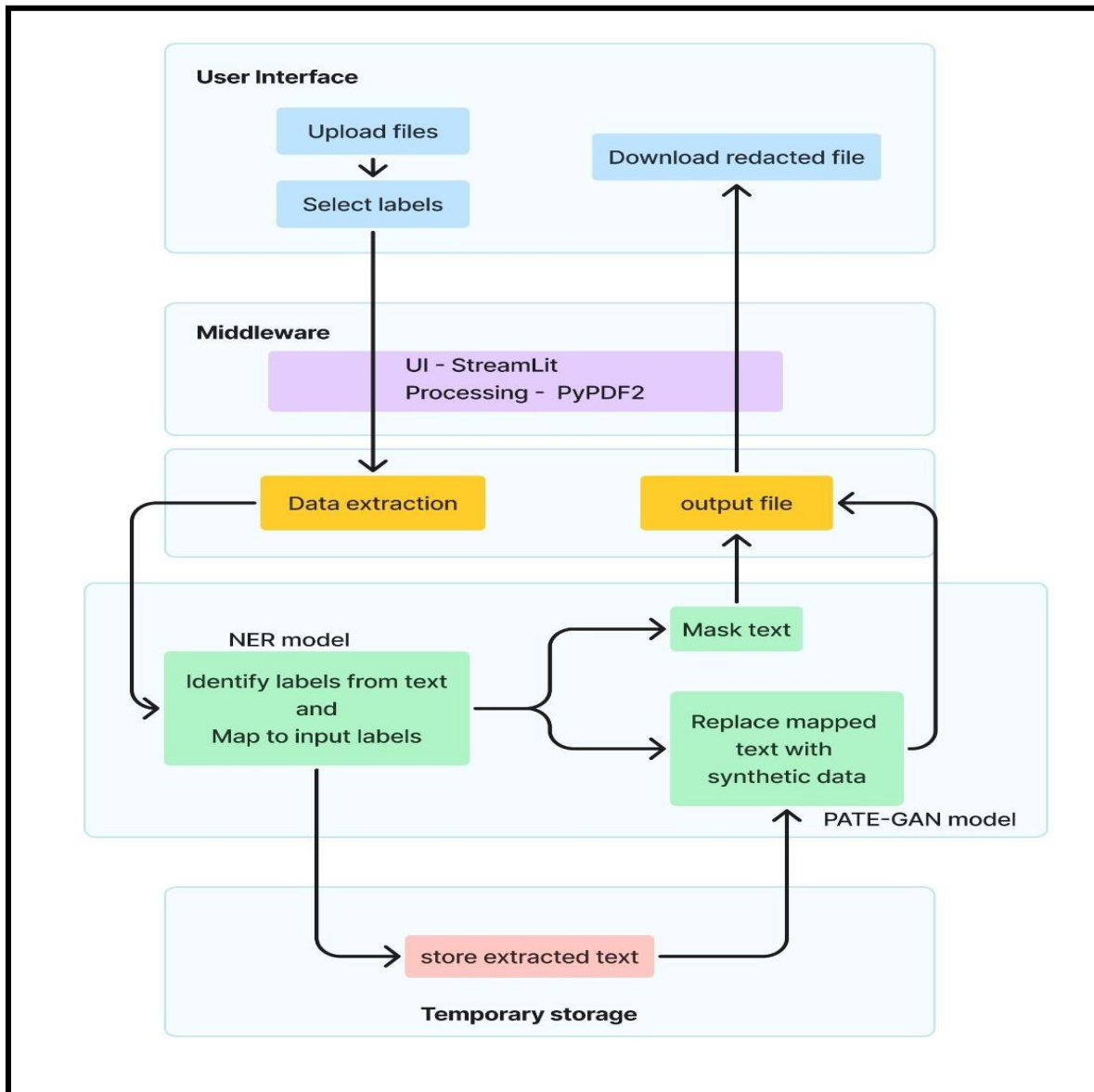


Fig. 3.1 Architecture Diagram of the Proposed Work

3.1 NER Model

This diagram(Fig 3.2) shows the training process of a machine learning model. It starts with training data that is divided into text and labels. These are used to calculate gradients, which help to improve the model. The model is then updated using these gradients and saved. The updated model makes predictions, which are compared with the actual labels to further refine and improve the model in an ongoing cycle.

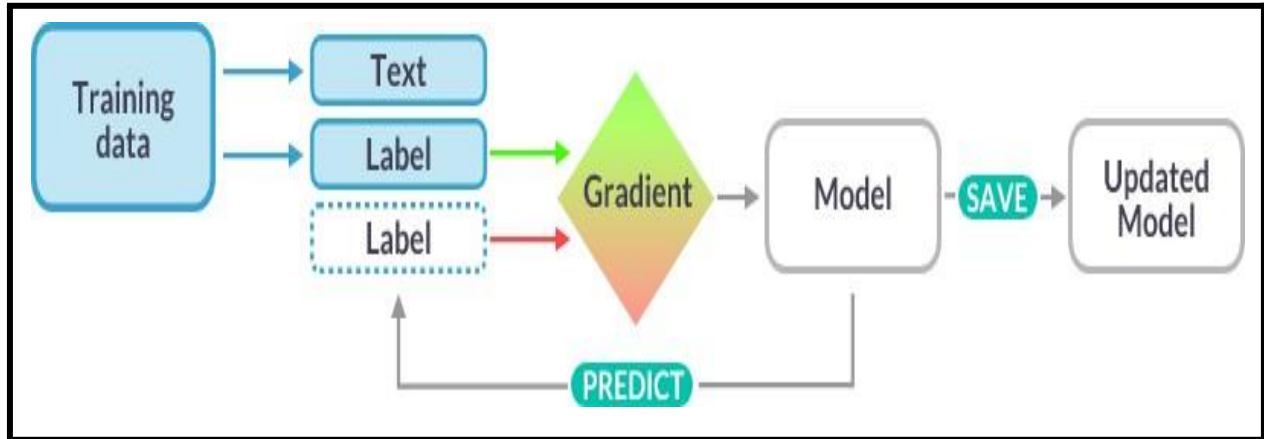


Fig. 3.2 NER Architecture

3.1.1 Data Collection

The AI4Privacy dataset was used to train the Named Entity Recognition (NER) model for identifying personally identifiable information (PII) in text. The dataset is synthetic, containing a variety of PII types across multiple languages. It includes 406,896 entries, with 17 PII classes. For the purpose of our redaction project, we focused only on English-language entries to streamline the training process and ensure the model's relevance for English text.

- Languages: English (85,321 entries), Italian, French, German, Dutch, and Spanish.
- PII Classes: 17 core labels like USERNAME, EMAIL, PHONE, ADDRESS, etc.

3.1.2 Data Extraction and Preprocessing

Language Filtering: Rows with non-English entries were removed, leaving only 85,321 English text samples. This ensured that the model focused on the target language, reducing complexity and improving accuracy.

Data Cleaning: Columns that were unnecessary, such as those related to languages and locales, were dropped. Each entry in the dataset contains text along with PII labels, such as USERNAME, EMAIL, and TIME, which were preserved for model training.

3.1.3 Tokenization and Labelling

The tokenization process involved breaking down the text into individual tokens, using BERT's tokenizer. The BIO tagging scheme was applied to label tokens as belonging to specific PII categories

B-: Beginning of an entity (e.g., B-USERNAME)

I-: Inside of an entity (e.g., I-USERNAME)

O: Outside of any entity (e.g., regular text not part of a PII).

3.1.4 Data Splitting

The dataset was split into training and validation sets, with 80% allocated for training (68,257 entries) and 20% for validation (17,064 entries). This split was crucial for evaluating the model's performance on unseen data and tuning it for better accuracy.

Table 3.1 Specifications of NER Model

Attribute	Details
Model	Piranha v1, fine-tuned from microsoft/mdeberta-v3-base
Task	Personal Identifiable Information detection
Fine Tuned	The model is fine-tuned for token classification to identify various types of PII.
Training Data	Trained on unstructured text data, specifically designed to detect PII such as account names, banking details, personal information, and more.
Context Length	256 tokens (DeBERTa)
Use Case	PII detection in unstructured text like documents, emails, and user-generated content.
PII Types Supported	17 types, including account numbers, credit card info, emails, passwords, SSNs, and more.

Attribute	Details
Model	Piranha v1, fine-tuned from microsoft/mdeberta-v3-base
Task	Personal Identifiable Information detection
Accuracy	99.44%
Loss	0.0173
Metrics	Precision: 93.16%, Recall: 93.08%, F1: 93.12%

Table 3.2 Hyperparameters Tuning of NER Model

Hyperparameter	Specification
Learning Rate	5E-05
Train Batch Size	128
Eval Batch Size	128
Seed	42
Optimizer	Adam (betas=(0.9, 0.999), epsilon=1e-08)
LR Scheduler Type	Linear
Activation function	GeLu
Number of Epochs	5

3.1.5 Model Architecture

The DeBERTa-v2 model is composed of 12 hidden layers, each with 12 attention heads, and features an embedding size of 768 (hidden size). It utilizes a feedforward network with an intermediate size typically set to 3072 and employs the GELU (Gaussian Error Linear Unit) activation function. Layer normalization is applied with an epsilon value of 1e-7 for numerical stability. The model supports a maximum of 512 position embeddings and uses 256 position

buckets for enhanced positional encoding. Its vocabulary consists of 251,000 tokens, enabling it to handle a wide range of linguistic nuances.

3.1.6 Loss Function -formula

The DeBERTa model for token classification uses **cross-entropy loss**. This loss function is commonly employed in multi-class classification tasks, where the objective is to minimize the difference between predicted probabilities and actual class labels. During training, the model aims to reduce the cross-entropy loss, thereby improving its accuracy in predicting the correct entity class for each token in the input sequence.

$$L(y, \hat{y}) = -\frac{1}{N} \sum_i \sum_j^C y_{ij} \log(\hat{y}_{ij})$$

Equation : 1

3.2 PATE-GAN

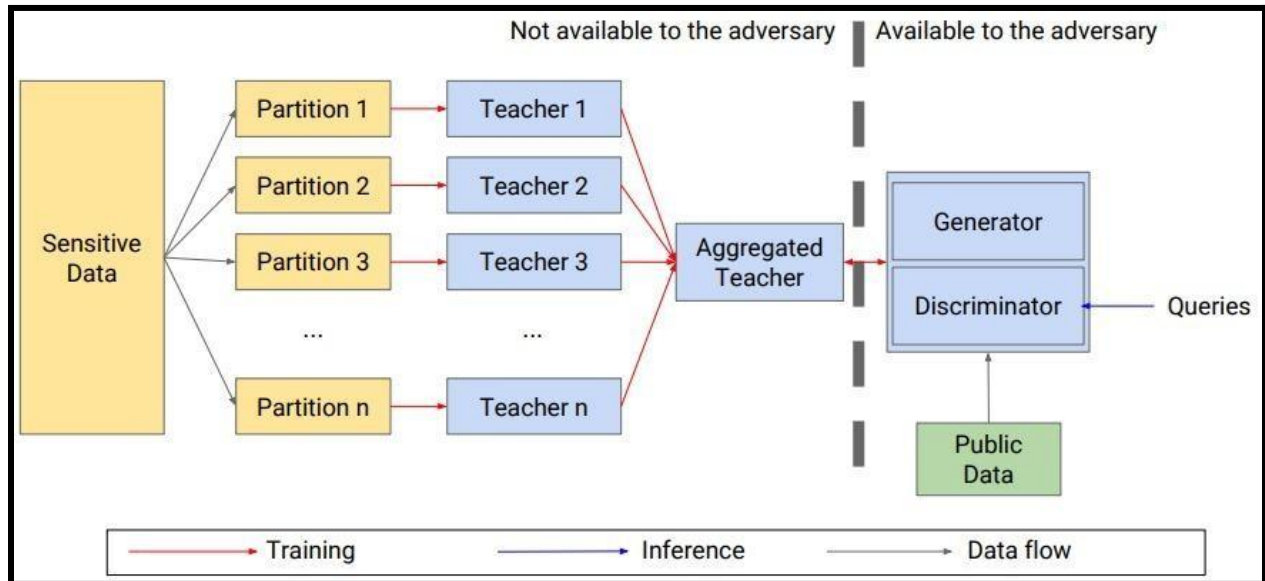


Fig 3.3 PATE Architecture for synthetic data generation

"Private Aggregation of Teacher Ensembles" (PATE) is a method used to protect privacy while training machine learning models. In this approach, several teacher models are trained on private

data, which is kept confidential. When a query is made, each teacher model provides its prediction, and these predictions are combined into a vote count. This count shows how many teachers agree on each prediction. The student model is then trained using public, unlabeled data and learns from these aggregated votes, rather than accessing the private data directly. This ensures that the privacy of the data is maintained, as the student model only learns from the aggregated predictions and not the private details.

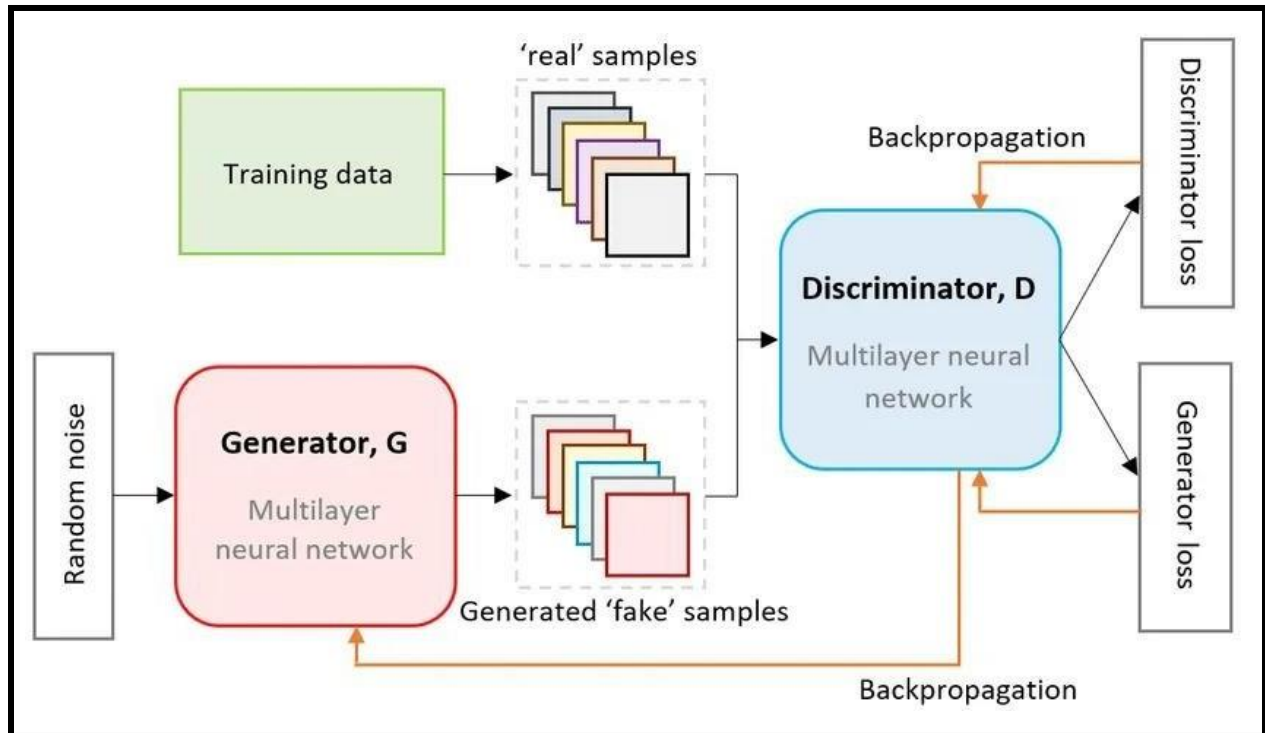


Fig 3.4 GAN Architecture for synthetic data generation

The diagram(Fig 3.4) illustrates the architecture of a Generative Adversarial Network (GAN), comprising two main components: the Generator (G) and the Discriminator (D). The Generator, a neural network, takes random noise as input and generates synthetic data samples ('fake' samples). The Discriminator, another neural network, evaluates and distinguishes between real samples (from the training data) and generated fake samples. The Discriminator's feedback guides the Generator through backpropagation, improving its ability to produce realistic data. This adversarial process continues, optimizing the Generator to produce data indistinguishable from real samples, while the Discriminator learns to improve its classification accuracy.

3.2.1 Data Collection

To train the PATE-GAN model for generating synthetic PII (Personally Identifiable Information), we created a custom synthetic dataset using the Faker library. The dataset mimics real-world PII data, such as emails, credit card numbers, names, and addresses, but all the data is entirely fake. This dataset serves as a foundation for training the PATE-GAN to generate privacy-preserving synthetic PII.

The steps for dataset generation are as follows:

We used the Faker library to generate fake PII data for several categories, including account numbers, building numbers, credit card numbers, and other personal identifiers like DOB, street addresses, and phone numbers. For example, fake names (given name, surname), email addresses, social security numbers, and more were randomly created. The synthetic dataset consists of labeled entries for various types of PII: I-ACCOUNTNUM, I-BUILDINGNUM, I-CITY, I-EMAIL, I-DATEOFBIRTH, etc. The data was structured into a pandas DataFrame, with each column corresponding to a type of PII entity.

3.2.2 Data Extraction and Preprocessing

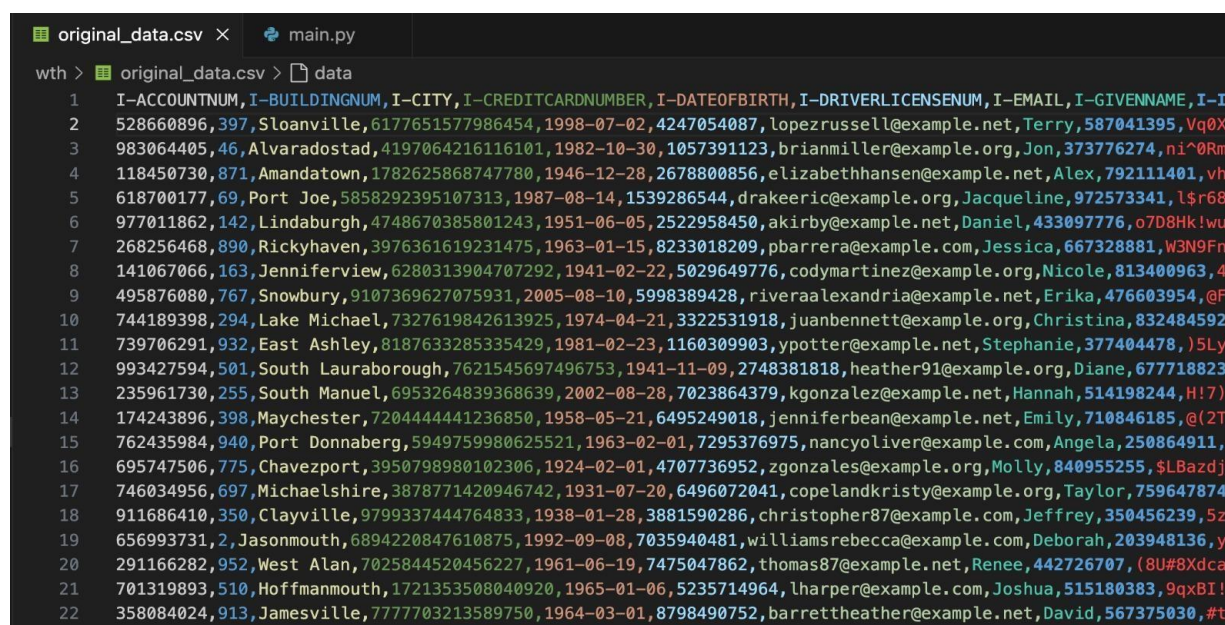
Categorical data was converted into numeric values using label encoding, and the entire dataset was normalized into the range $[0, 1]$ using MinMaxScaler. This ensures that the PATE-GAN model could effectively learn from the data. The dataset was split into 80% training data and 20% testing data using `train_test_split` from `sklearn.model_selection`. The PATE-GAN model was trained to generate synthetic PII data using an architecture consisting of a generator and a discriminator. The generator learns to create synthetic PII data by identifying patterns in the labeled dataset, while the discriminator is tasked with distinguishing between real and synthetic data. Adversarial training was employed to enhance the generator's ability to produce realistic synthetic data and refine the discriminator's capacity to detect synthetic inputs. To ensure privacy and prevent overfitting, differential privacy was incorporated, safeguarding against the model memorizing sensitive data.

The PATE-GAN framework, which uses multiple teacher models to aggregate their outputs and maintain privacy, was applied. This ensures that individual PII entities are not linked to any

particular real-world person. The training process was conducted with a batch size of 32, a learning rate of $1e-4$, and spanned 10 epochs. The loss function incorporated both adversarial loss and differential privacy loss, ensuring the generated synthetic data was realistic while maintaining privacy.

3.2.3 Synthetic Data Example

Here's an example of how synthetic PII data looks after training: (Fig 3.5 Dataset Example).



```

1 I-ACCOUNTNUM, I-BUILDINGNUM, I-CITY, I-CREDITCARDNUMBER, I-DATEOFBIRTH, I-DRIVERLICENSENUM, I-EMAIL, I-GIVENNAME, I-IR
2 528660896, 397, Sloanville, 6177651577986454, 1998-07-02, 4247054087, lopezrussell@example.net, Terry, 587041395, Vq0X
3 983064405, 46, Alvaradostad, 4197064216116101, 1982-10-30, 1057391123, brianmiller@example.org, Jon, 373776274, ni^0Rmi
4 118450730, 871, Amandatown, 1782625868747780, 1946-12-28, 2678800856, elizabethhansen@example.net, Alex, 792111401, vhu
5 618700177, 69, Port Joe, 5858292395107313, 1987-08-14, 1539286544, drakeeric@example.org, Jacqueline, 972573341, l$R68J
6 977011862, 142, Lindaburgh, 4748670385801243, 1951-06-05, 2522958450, akirby@example.net, Daniel, 433097776, o7D8Hk!wu0
7 268256468, 890, Rickyhaven, 3976361619231475, 1963-01-15, 8233018209, pbarrera@example.com, Jessica, 667328881, W3N9Fn
8 141067066, 163, Jenniferview, 6280313904707292, 1941-02-22, 5029649776, codymartinez@example.org, Nicole, 813400963, 4c
9 495876080, 767, Snowbury, 9107369627075931, 2005-08-10, 5998389428, riveraaalexandria@example.net, Erika, 476603954, @FS
10 744189398, 294, Lake Michael, 7327619842613925, 1974-04-21, 3322531918, juanbennett@example.org, Christina, 832484592,
11 739706291, 932, East Ashley, 8187633285335429, 1981-02-23, 1160309903, ypotter@example.net, Stephanie, 377404478, )5Lyf
12 993427594, 501, South Lauraborough, 7621545697496753, 1941-11-09, 2748381818, heather91@example.org, Diane, 677718823,
13 235961730, 255, South Manuel, 6953264839368639, 2002-08-28, 7023864379, kgonzalez@example.net, Hannah, 514198244, H!7)9
14 174243896, 398, Maychester, 7204444441236850, 1958-05-21, 6495249018, jenniferbean@example.net, Emily, 710846185, @2Th
15 762435984, 940, Port Donnaberg, 5949759980625521, 1963-02-01, 7295376975, nancyoliver@example.com, Angela, 250864911,
16 695747506, 775, Chavezport, 3950798980102306, 1924-02-01, 4707736952, zgonzaless@example.org, Molly, 840955255, $LBazdj6
17 746034956, 697, Michaelshire, 3878771420946742, 1931-07-20, 6496072041, copelandkristy@example.org, Taylor, 759647874,
18 911686410, 350, Clayville, 9799337444764833, 1938-01-28, 3881590286, christopher87@example.com, Jeffrey, 350456239, 5zC
19 656993731, 2, Jasonmouth, 6894220847610875, 1992-09-08, 7035940481, williamsrebecca@example.com, Deborah, 203948136, yf
20 291166282, 952, West Alan, 7025844520456227, 1961-06-19, 7475047862, thomas87@example.net, Renee, 442726707, (8U#8Xdca/
21 701319893, 510, Hoffmanmouth, 1721353508040920, 1965-01-06, 5235714964, lharper@example.com, Joshua, 515180383, 9qxBI!1
22 358084024, 913, Jamesville, 7777703213589750, 1964-03-01, 8798490752, barrettheather@example.net, David, 567375030, #t

```

Fig 3.5 Dataset Example for PATE-GAN

3.2.3 Integration with NER Model

The synthetic PII data generated by PATE-GAN is used to replace the real PII in the documents identified by the NER (Named Entity Recognition) model. This ensures that sensitive information is redacted while maintaining the integrity and structure of the original text.

By generating a custom synthetic dataset using Faker, we were able to create a diverse set of labeled PII data to train the PATE-GAN model. The model, in turn, was able to learn how to generate realistic synthetic PII that closely mimics real-world data, while ensuring privacy through differential privacy measures.

CHAPTER 4

RESULTS AND DISCUSSION

In our project, we used the TPU v2-8 from Google Cloud for its powerful computing capabilities, along with the PyTorch and TensorFlow frameworks for building and training models. We also used the Hugging Face Transformers library for NLP tasks with pre-trained models like BERT and DeBERTa. This combination helped us handle large datasets and complex neural networks efficiently, reducing training time and improving performance.

The TPU v2-8 is a Google Cloud Tensor Processing Unit with eight cores, providing strong computational power for deep learning. PyTorch is a popular framework for building and training models, while TensorFlow is widely used for both research and production. The Hugging Face Transformers library is excellent for NLP tasks, especially with pre-trained transformer models. Together, these tools offer a powerful suite for advanced machine learning development.

4.1 Qualitative Analysis

The user interface for the redact tool, helps users to upload files, select entities for redaction or opt for automatic PII Redaction. Redacted document preserves the post redaction format and users can upload files across various formats like text files(.txt), pdfs(.pdf) and word documents(.docx). They can also download the redacted documents.

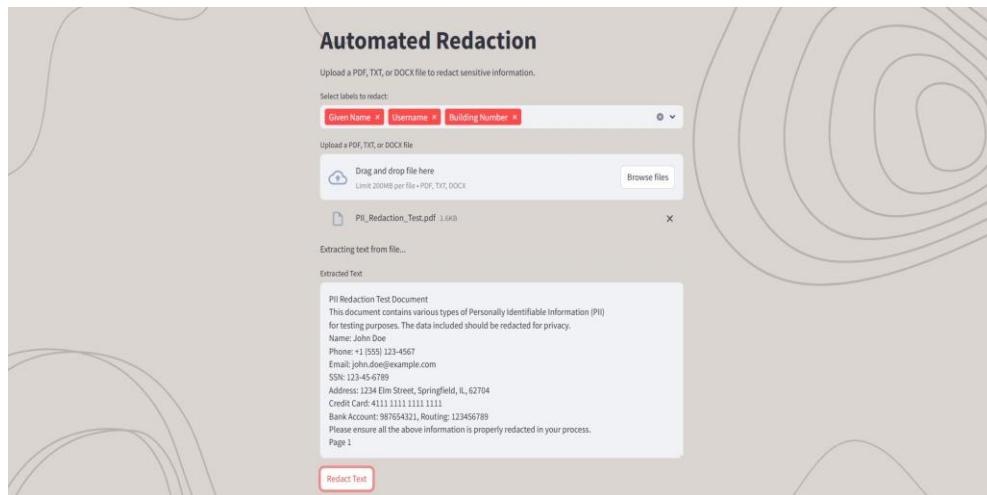


Fig. 4.1 User Interface

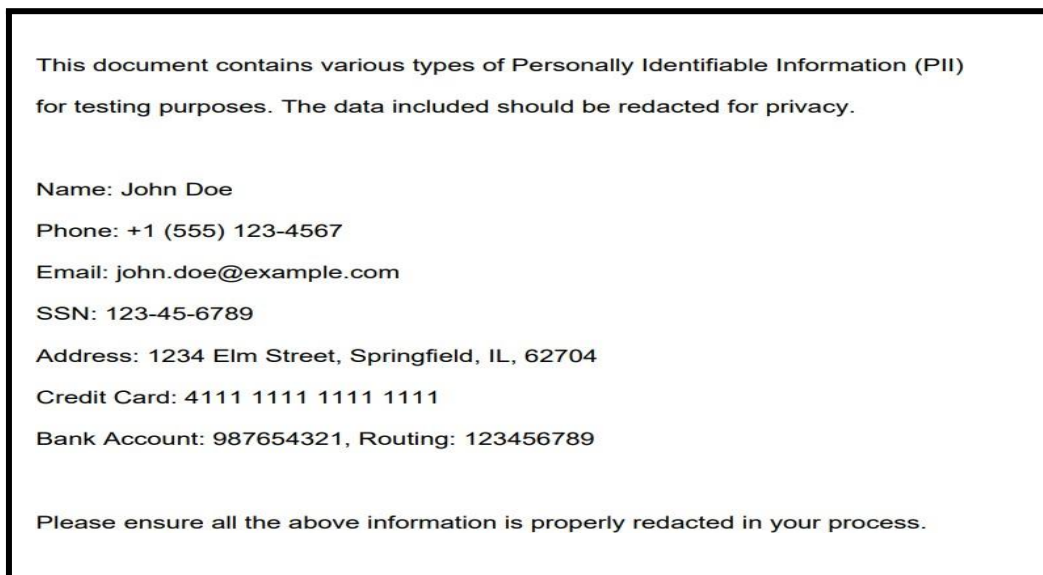


Fig. 4.2 Original document

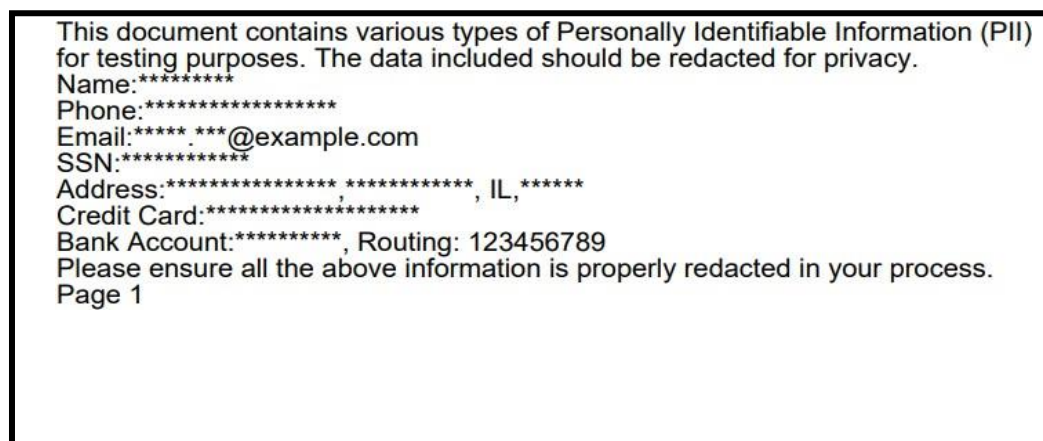


Fig. 4.3 Redacted document

4.2 Quantitative Analysis

To check the performance of NER, we have mentioned the following metrics:

F1-Score: Comprehensive evaluation combining precision and recall for each entity type and overall.

Precision: Proportion of correctly redacted entities out of all detected entities.

Recall: Proportion of correctly redacted entities out of all true sensitive entities.

Efficiency:

Runtime per document and overall system throughput in terms of redacted files per second.

Evaluation of memory and computational resource usage on hardware (e.g., TPU v2-8).

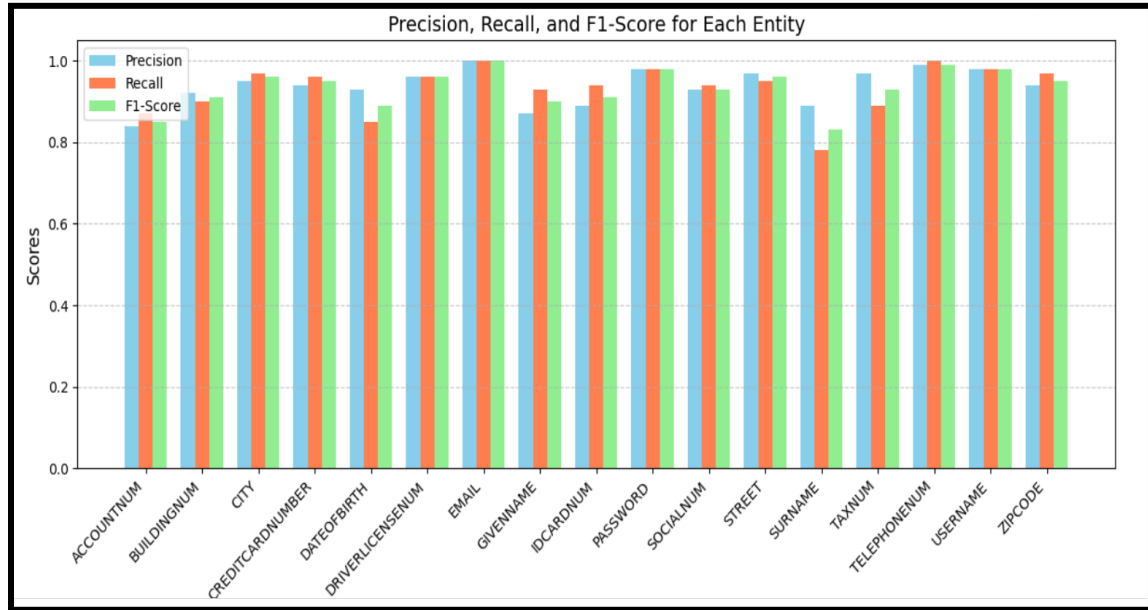


Fig 4.4 Precision, Recall, and F-1 Score for Each Entity

The above bar graph(Fig 4.4) evaluates the performance of an entity recognition system by analyzing Precision, Recall, and F1-Score across various entity types, including Account Number, Building Number, City, and others. Most entities achieve consistently high scores, close to 1, indicating accurate and reliable detection. However, entities like Building Number and Driver License Number exhibit slightly lower performance, suggesting areas for improvement.

The close alignment of Precision, Recall, and F1-Score across all entities demonstrates a balanced performance with minimal false positives and negatives. This analysis highlights the system's robustness while identifying specific entities that may benefit from further optimization.

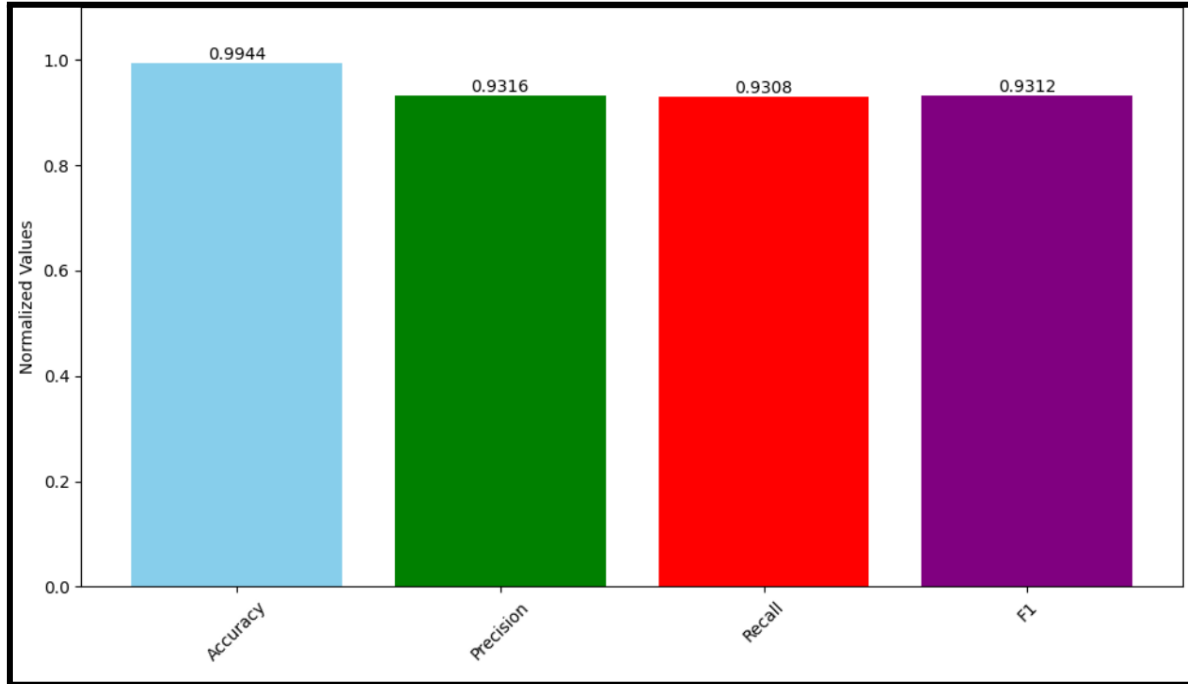


Fig 4.5 Performance metrics for NER

The bar chart mentioned above (Fig 4.5) illustrates the performance metrics of a Named Entity Recognition (NER) model. The model achieves high accuracy (0.9944), indicating strong overall correctness, and balanced precision (0.9316), recall (0.9308), and F1 score (0.9312), reflecting its effectiveness in identifying and classifying entities accurately and consistently.

To check the performance of PATE-GAN, we have mentioned the following metrics:

data_no: Number of generated data.

data_dim: Number of dimensions of generated data (if random).

dataset: Dataset to use.

noise_rate: Noise ratio on data.

iterations: Number of iterations for handling randomness.

n_s: Number of student training iterations.

batch_size: Batch size for training student and generator.

k: Number of teachers.

epsilon: Differential privacy parameter (epsilon).

delta: Differential privacy parameter (delta).

lambda: PATE noise size.

Table 4.1 Parameters and Default Values

Best-Perf:0.8136282761471684

Parameter	Default Value
data_no	10000
data_dim	10
dataset	Generated with faker
noise_rate	1.0
iterations	50
n_s	1
batch_size	64
k	10
epsilon	1.0
delta	0.00001
lamda	1.0

```
Averages:
AUC-Original      0.829155
AUC-Synthetic     0.739873
APR-Original      0.870227
APR-Synthetic     0.796636
dtype: float64
```

Fig 4.6 Averages for Default Values

Table 4.2 Privacy-Heavy Configuration**Best-Perf:0.800200032324599**

Parameter	Value
data_no	10000
data_dim	10
dataset	random
noise_rate	1.0
iterations	50
n_s	1
batch_size	32
k	5
epsilon	0.1
delta	0.000001
lamda	2.0

```

Averages:
AUC-Original      0.813791
AUC-Synthetic     0.764191
APR-Original      0.821082
APR-Synthetic     0.760391
dtype: float64

```

Fig 4.7 Averages for Privacy-Heavy Configuration**Table 4.3** Utility-Heavy Configuration:**Best-Perf:0.8251600150092399**

Parameter	Value
data_no	10000

data_dim	10
dataset	random
noise_rate	1.0
iterations	50
n_s	1
batch_size	128
k	20
epsilon	2.0
delta	0.001
lamda	0.5

```

Averages:
AUC-Original      0.848900
AUC-Synthetic     0.775918
APR-Original      0.874345
APR-Synthetic     0.799345
dtype: float64

```

Fig 4.8 Averages for Utility-Heavy Configuration

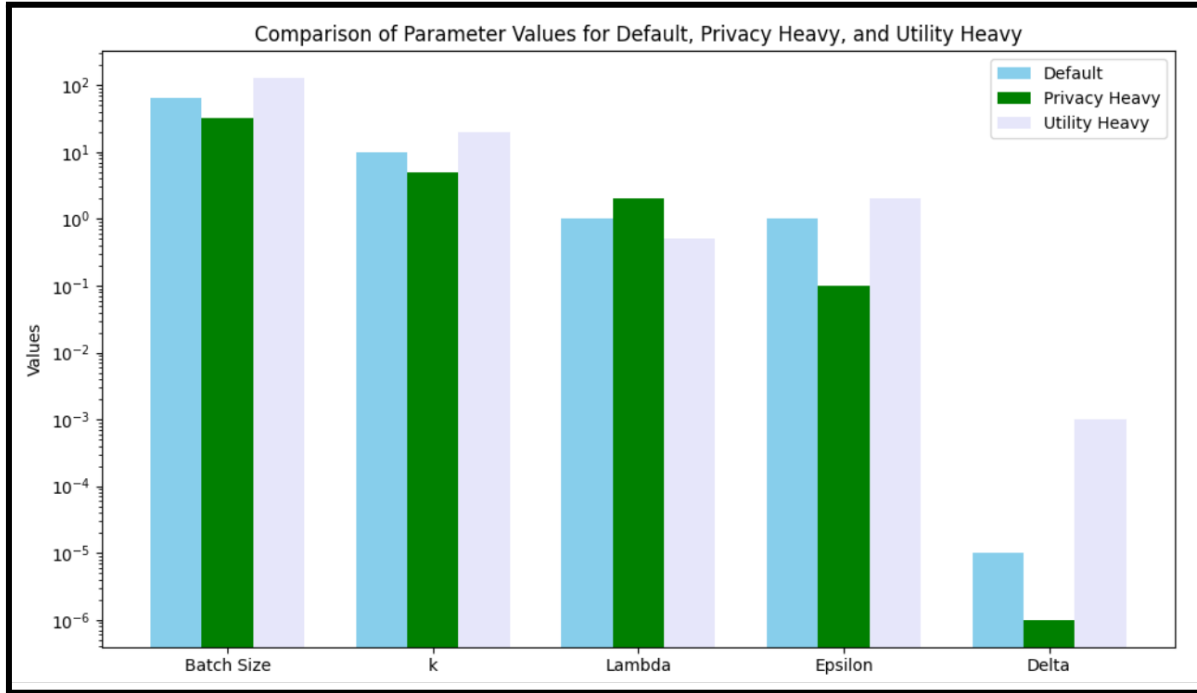


Fig 4.9 Comparison of Parameter Values

The above bar graph (Fig 4.9) presents a logarithmic-scale comparison of key parameter values—Batch Size, k, Lambda, Epsilon, and Delta—across Default, Privacy Heavy, and Utility Heavy configurations. Privacy Heavy settings exhibit lower values for Epsilon and Delta, reflecting tighter differential privacy constraints, while Utility Heavy configurations maximize utility with elevated k and Delta values, indicative of relaxed privacy constraints. Default configurations adopt intermediate parameter values, balancing the trade-offs between privacy preservation and utility optimization. This analysis underscores the impact of parameter tuning on privacy-utility trade-offs in model configurations.

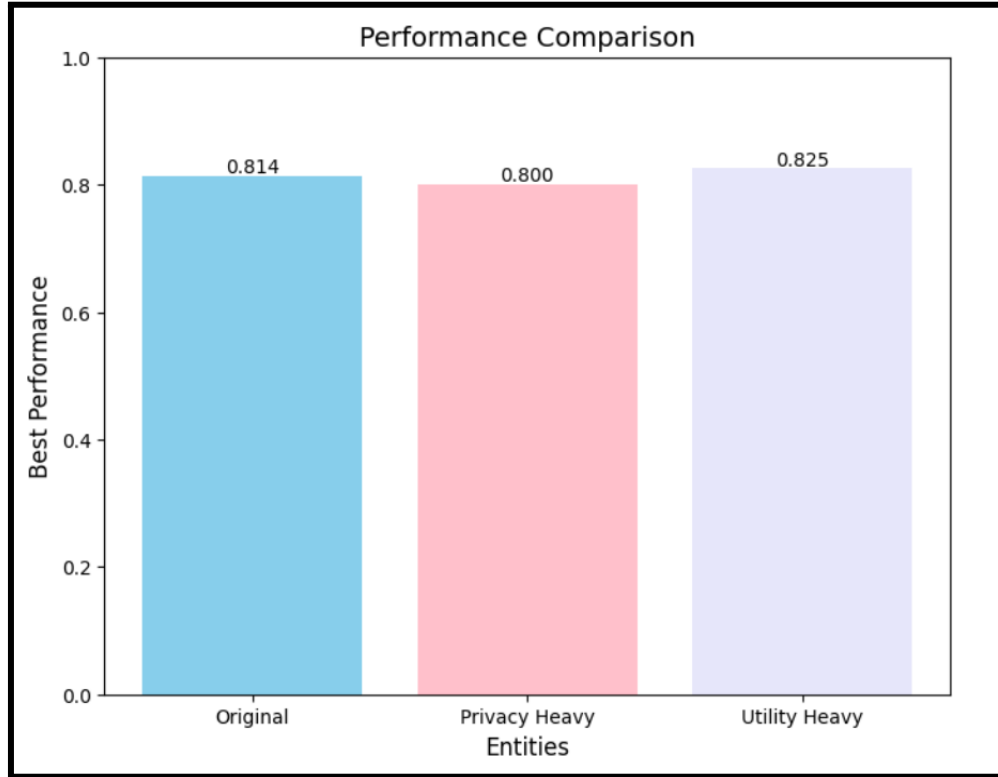


Fig 4.10 Performance Comparison of Original, Privacy Heavy and Utility Heavy

Among the three configurations tested (default, privacy-heavy, and utility-heavy)(Fig 4.10), the utility-heavy configuration achieved the highest performance, with a Best-Perf score of 0.8252. This superior performance highlights the trade-off between utility and privacy, as the utility-heavy setup prioritized model accuracy by increasing parameters like batch size and differential privacy thresholds (epsilon and lambda). While this approach resulted in longer execution times, it demonstrates that relaxing privacy constraints can significantly enhance model performance when utility is the primary objective.

CHAPTER 5

CONCLUSION AND FUTURE SCOPE

5.1 Conclusion

The proposed work successfully integrates Named Entity Recognition (NER) and PATE-GAN models to provide a comprehensive automated redaction tool for sensitive data. By leveraging the AI4Privacy dataset and generating synthetic datasets using the Faker library, the system ensures high accuracy in detecting and anonymizing personally identifiable information (PII). The NER model achieved an impressive accuracy of 99.44%, while the PATE-GAN model effectively with an accuracy of 81.363% while generating new synthetic data and also preserving the contextual and structural integrity of the original documents. Preprocessing steps, including language filtering, tokenization, and normalization, contributed to the robust performance of the models. This tool addresses critical challenges in data privacy by providing a scalable, efficient, and user-friendly solution suitable for industries like finance, and legal sectors.

5.2 Future Scope

The proposed system can be extended in several directions to enhance its functionality and applicability. **Language expansion** is a key area, aiming to support multiple languages and cater to a global user base. Another focus is on **real-time redaction**, enabling the system to handle streaming data and dynamic text sources effectively. The user interface can be further improved by incorporating **advanced customization options** tailored to specific redaction needs. Enhancements in **synthetic data generation** techniques will improve the realism and adaptability of the generated data. Additionally, **cloud integration** can be implemented to process large-scale datasets efficiently and improve accessibility. Finally, aligning the tool with global data privacy laws, such as **GDPR** and **CCPA**, will ensure regulatory compliance across different jurisdictions.

References

- [1] Cumby, C., & Ghani, R. (2011, August). 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).

- [2] Yermilov, O., Raheja, V., & Chernodub, A. (2023). Privacy-and utility-preserving nlp with anonymized data: A case study of pseudonymization. *arXiv preprint arXiv:2306.05561*.

- [3] Peng, S., Huang, M. J., Wu, M., & Wei, J. (2024). Transforming Redaction: How AI is Revolutionizing Data Protection. *arXiv preprint arXiv:2409.15308*.

- [4] Lison, P., Pilán, I., Sánchez, D., Batet, M., & Øvrelid, L. (2021, August). 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).

- [5] Larbi, I. B. C., Burchardt, A., & Roller, R. (2022). Which anonymization technique is best for which NLP task?--It depends. A Systematic Study on Clinical Text Processing. *arXiv preprint arXiv:2209.00262*.

- [6] Ganev, G., Annamalai, M. S. M. S., & De Cristofaro, E. (2024). The Elusive Pursuit of Replicating PATE-GAN: Benchmarking, Auditing, Debugging. *arXiv preprint arXiv:2406.13985*.

- [7] Raj, A., & D'Souza, R. (2021). Anonymization of sensitive data in unstructured documents using NLP. *International Journal of Mechanical Engineering and Technology (IJMET)*, 12(4), 25-35.

- [8] Stolfo, S. J., Wolfson, O., Chan, P. K., Dewan, H. M., Woodbury, L., Glazier, J. S., & Ohsie, D. A. (1991). PARULEL: Parallel rule processing using meta-rules for redaction. *Journal of Parallel and Distributed Computing*, 13(4), 366-382.
- [9] Patsakis, C., & Lykousas, N. (2023). Man vs the machine in the struggle for effective text anonymisation in the age of large language models. *Scientific Reports*, 13(1), 16026.
- [10] Ma, Y., Mao, Q., & Hu, N. (2024). To protect or to hide-an investigation on corporate redacted disclosure motives under new FAST act regulation. *Emerging Markets Review*, 60, 101144.
- [11] Saxena, A. K. (2022). Enhancing Data Anonymization: A Semantic K-Anonymity Framework with ML and NLP Integration. *Sage Science Review of Applied Machine Learning*, 5(2), 81-92.
- [12] Gusain, V., & Leith, D. (2023, March). Towards Quantifying the Privacy of Redacted Text. In *European Conference on Information Retrieval* (pp. 423-429). Cham: Springer Nature Switzerland.
- [13] Felps, D. L., Schwickerath, A. D., Williams, J. D., Vuong, T. N., Briggs, A., Hunt, M., ... & Shumaker, T. (2020). Class clown: Data redaction in machine unlearning at enterprise scale. *arXiv preprint arXiv:2012.04699*.
- [14] Kong, Z., & Chaudhuri, K. (2023, February). Data redaction from pre-trained gans. In *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)* (pp. 638-677). IEEE.
- [15] Brennan, M., Afroz, S., & Greenstadt, R. (2012). Adversarial stylometry: Circumventing authorship recognition to preserve privacy and anonymity. *ACM Transactions on Information and System Security (TISSEC)*, 15(3), 1-22.