International Journal of Engineering Research and Science & Technology

AI-BASED PRIVACY PRESERVING TEXT TRANSFORMATION

GUIDE

[1]Name : Mrs.G.ANITHA

Assistant Professor

[2] SHIVANI CHIGULLAPALLY - cpshivani2004@**gmail.com**

[3] CH.BHARGAVI - bhargavich088@**gmail.com**

[4] C.SRIYA REDDY - csriyareddy1429@**gmail.com**

[5] G.SRUTHI - sruthigopagolla@**gmail.com**

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
VIGNAN'S INSTITUTE OF MANAGEMENT AND TECHNOLOGY FOR WOMEN
(An Autonomous Institution)
(Affiliated to Jawaharlal Nehru Technological University Hyderabad, Accredited by NBA,NAAC with A+)
Kondapur(Village), Ghatkesar (Mandal), Medchal (Dist.)
Telangana-501301

(2025-2026)

ABSTRACT

he rapid growth of digital platforms has significantly increased the risk of exposing sensitive personal information present in textual data. Protecting such data while maintaining its usability has become a critical challenge in modern data processing systems. This project, *"AI-Based Privacy-Preserving Text Transformation,"* presents an intelligent and automated solution for safeguarding sensitive information without compromising the original meaning and readability of the text.

The proposed system leverages advanced **Natural Language Processing (NLP)** techniques and machine learning tools such as spaCy and Microsoft Presidio to identify Personally Identifiable Information (PII), including names, phone numbers, email addresses, and other confidential data. Once detected, the system applies privacy-preserving techniques like masking, anonymization, and placeholder

substitution to secure the information while preserving contextual integrity.

Additionally, the system supports multi-format data processing by integrating **Optical Character Recognition (OCR)** technologies such as PyTesseract and PDF2Image, enabling text extraction from images and scanned documents. To ensure data security, the original sensitive information is encrypted using AES-based cryptographic methods and stored securely, allowing controlled access and restoration only by authorized users through a role-based system.

The architecture of the system, as illustrated in the *system design diagram on page 16*, follows a multi-layered approach involving preprocessing, analysis, transformation, validation, and secure storage, ensuring scalability and efficiency. The implementation demonstrates high accuracy in detecting sensitive entities and reliable restoration of original data, achieving a balance between data privacy and usability.

Keywords

AI-based Privacy Preservation, Text Anonymization, Personally Identifiable Information (PII), Natural Language Processing (NLP), Named Entity Recognition (NER), Data Masking, Microsoft Presidio, spaCy, Optical Character Recognition (OCR), Data Encryption, AES Security, Secure Data Storage, Role-Based Access Control (RBAC), Text Transformation, Data Privacy.

I. INTRODUCTION

In today's digital era, vast amounts of textual data are generated and shared across platforms such as social media, healthcare systems, financial services, and enterprise applications. This data often contains sensitive information, including names, contact details, addresses, and other Personally Identifiable Information (PII). With the increasing dependence on digital communication, the risk of data breaches, identity theft, and unauthorized access to confidential information has grown significantly. Protecting such sensitive data while maintaining its usability for analysis and communication has become a major challenge.

Traditional approaches to data privacy, such as manual redaction, rule-based masking, and encryption, have several limitations. Manual methods are time-consuming and prone to human error, while rule-based systems lack contextual understanding and fail to detect complex or unstructured sensitive information.

Although encryption ensures data security, it restricts direct usability of the data without decryption, making it less suitable for real-time processing and analysis. As highlighted in the document, these methods often fail to provide a balance between **data privacy and data utility**.

To address these challenges, Artificial Intelligence (AI) and Natural Language Processing (NLP) have emerged as powerful solutions for automated privacy preservation. NLP techniques, particularly Named Entity Recognition (NER), enable systems to identify sensitive entities within unstructured text with high accuracy. By combining machine learning with intelligent text processing, it is possible to automate the detection and transformation of private data while preserving the original meaning and readability of the content.

This project, *"AI-Based Privacy-Preserving Text Transformation,"* proposes an intelligent system that automatically detects and protects sensitive information in textual data. The system utilizes advanced NLP tools such as spaCy and Microsoft Presidio for entity detection, along with anonymization techniques like masking and placeholder substitution to ensure privacy. Additionally, it integrates Optical Character Recognition (OCR) technologies to process text from images and scanned documents, enabling multi-format support.

II. LITERATURE REVIEW

The problem of protecting sensitive information in textual data has been widely studied in the fields of **Natural Language Processing (NLP)** and data privacy. Several research works and systems have focused on detecting and anonymizing Personally Identifiable Information (PII) using different techniques, ranging from rule-based approaches to advanced deep learning models.

Early approaches to privacy preservation relied on **rule-based systems and regular expressions** to identify structured data such as phone numbers, email addresses, and identification numbers. While these methods are simple and computationally efficient, they lack contextual understanding and often fail to detect unstructured or complex entities. As a result, their accuracy is limited, especially when dealing with large and diverse datasets.

With the advancement of NLP, **Named Entity Recognition (NER)** models have been widely adopted for detecting sensitive entities in text. Research by Lample et al. (2016) introduced neural architectures for NER, demonstrating significant

improvements in entity recognition using deep learning techniques. Similarly, Meystre et al. (2010) focused on de-identification of medical records, highlighting the importance of automated systems in protecting patient data while maintaining usability.

Recent studies have incorporated **machine learning and hybrid approaches** for improved performance. As shown in the *literature survey table on page 6* , models such as BiLSTM, BERT, and spaCy-based systems have been used to enhance detection accuracy. For instance, Dias et al. (2020) achieved high accuracy using hybrid NER models combined with rule-based methods, while Szwernia et al. (2024) utilized KB-BERT models for PII detection in textual data. These approaches demonstrate better contextual understanding compared to traditional methods. In addition, modern systems like **Microsoft Presidio** integrate multiple techniques, including NER, pattern recognition, and context-based analysis, to identify sensitive data across various formats. Research also highlights the use of **OCR technologies**, such as Tesseract, to extract text from images and scanned documents, enabling privacy preservation in multi-format data sources.

Some studies have explored **privacy in Large Language Models (LLMs)**, focusing on adversarial desensitization and context-aware anonymization. While these approaches improve flexibility and detection capabilities, they often involve high computational costs and complexity. Other research has emphasized improving precision and recall through hybrid systems that combine statistical models with deterministic rules.

III. METHODOLOGY

1. Data Input and Preprocessing

The process begins with user input, where the system accepts text or files in multiple formats such as PDF, DOCX, TXT, and images. For scanned documents or images, OCR tools like PyTesseract are used to extract textual content. The extracted text is then preprocessed by removing noise, normalizing formatting, and preparing it for analysis.

2. Sensitive Data Detection (PII Identification)

The system uses a hybrid detection approach to identify sensitive information:

- **Named Entity Recognition (NER):** NLP models (spaCy and

Presidio) detect entities such as names, locations, and organizations.

- **Pattern Matching (Regex):** Used to identify structured data like phone numbers, Aadhaar numbers, PAN, and email IDs.
- **Score-Based Filtering:** Each detected entity is assigned a confidence score, and only high-confidence entities are selected for processing.

This combination improves accuracy and ensures both structured and unstructured data are identified effectively.

3. Conflict Resolution Mechanism

In cases where multiple entities overlap (e.g., a number detected as both phone and ID), a conflict resolution algorithm is applied. The system:

- Sorts detected entities based on position and confidence
- Retains the most relevant entity
- Eliminates overlaps to avoid incorrect masking

This ensures clean and precise anonymization without data corruption.

4. Data Anonymization and Transformation

Once sensitive entities are identified, the system replaces them using privacy-preserving techniques:

- **Masking** (e.g., XXXXX1234)
- **Placeholder substitution** (e.g., <NAME_1>)
- **Data generalization**

The transformation is performed carefully to preserve the original meaning and readability of the text.

5. Secure Data Storage (Encryption Layer)

To ensure security, the original sensitive data is not stored in plain text. Instead:

- Data is encrypted using **AES-128 (Fernet encryption)**
- A mapping is maintained between placeholders and encrypted original values
- Data is stored securely in a database (MySQL)

This "vault-based" approach ensures that even if the database is compromised, sensitive data remains protected.

6. Role-Based Access and Data Restoration

The system implements **Role-Based Access Control (RBAC)**:

- **Standard Users:** Can upload and anonymize data
- **Administrators:** Can restore original data when required

During restoration:

- Encrypted values are decrypted
- Placeholders are replaced with original data
- The document is reconstructed accurately without loss of context

7. Multi-Format Processing with OCR Support

To enhance usability, the system supports:

- Text documents (TXT, DOCX, PDF)
- Scanned files and images via OCR

If direct text extraction fails, OCR is used as a fallback to ensure all content is processed effectively.

8. System Integration and Output Generation

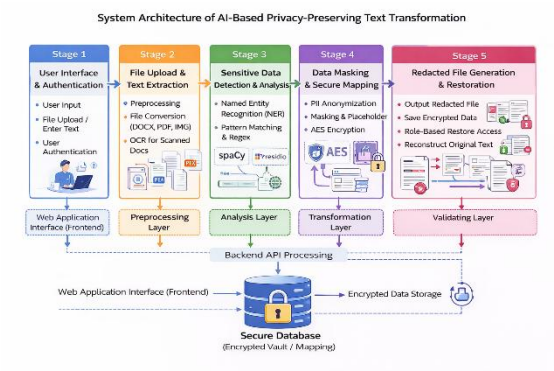
The entire system is implemented as a web application:

- **Frontend:** React (user interface)
- **Backend:** Flask (API processing)
- **Database:** MySQL (secure storage)

The final output provided to the user includes:

- Anonymized text/document
- Unique reference ID for tracking and restoration

IV. SYSTEM ARCHITECTURE



V. RESULTS & DISCUSSION

The proposed AI-Based Privacy-Preserving Text Transformation system was evaluated for its effectiveness in detecting, anonymizing, and securely restoring sensitive information in textual data. The results indicate that the system achieves high accuracy in identifying Personally Identifiable Information (PII) by combining Natural Language Processing (NLP) techniques with pattern-based detection methods. Entities such as names, phone numbers, email addresses, and identification numbers were successfully recognized with minimal false positives, demonstrating the strength of the hybrid detection approach.

The anonymization process proved to be efficient and reliable, as sensitive data was replaced with placeholders or masked values while preserving the overall meaning and readability of the text. This ensures that the transformed data remains useful for analysis and communication, addressing the critical challenge of balancing data privacy with data utility. The system also maintained structural consistency in the output, which is essential for practical applications across various domains.

Performance evaluation shows that the system operates with high efficiency. Text-based anonymization is completed in less than 500 milliseconds, while OCR-based processing for documents and images takes approximately 2.5 to 5 seconds depending on the complexity of the input. The system also demonstrated strong stability under concurrent processing conditions, handling multiple requests without failures. Additionally, the encryption mechanism ensures secure storage of original data, and the restoration process achieved 100% reliability with negligible latency.

From a discussion perspective, the system significantly improves upon traditional privacy-preserving methods such as manual redaction and rule-based masking. The use of NLP enables contextual understanding, which enhances detection accuracy compared to conventional techniques. The integration of OCR further extends the system's capability to process non-digital and scanned documents, making it more versatile for real-world scenarios.

VI. CONCLUSION

The *AI-Based Privacy-Preserving Text Transformation* system successfully addresses the growing need for protecting sensitive information in textual data while maintaining its usability. In an era where data breaches and privacy concerns are increasing, the proposed solution provides an effective and intelligent approach to safeguard Personally Identifiable Information (PII) across various domains.

The system integrates advanced technologies such as Natural Language Processing (NLP), Named Entity Recognition (NER), pattern matching, and Optical Character Recognition (OCR) to accurately detect sensitive data from both structured and unstructured sources. By applying anonymization techniques like masking and placeholder substitution, it ensures that confidential information is protected without affecting the readability and context of the text.

A key strength of the system lies in its secure architecture, where original data is

encrypted using AES-based cryptographic methods and stored safely. The implementation of Role-Based Access Control (RBAC) further enhances security by allowing only authorized users to restore original data when required. This ensures both privacy and controlled accessibility, which is essential in real-world applications.

The system demonstrates high accuracy, efficiency, and reliability, as validated through testing. It performs well across multiple file formats, including text documents and images, making it a flexible and scalable solution. Compared to traditional methods, the proposed approach offers improved automation, reduced human error, and better context preservation.

REFERENCES

1. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). *Neural Architectures for Named Entity Recognition*. Proceedings of NAACL-HLT.
2. Meystre, S. M., Friedlin, F. J., South, B. R., Shen, S., & Samore, M. H. (2010). *Automatic De-identification of Electronic Medical Records*. Journal of the American Medical Informatics Association.
3. Dernoncourt, F., Lee, J. Y., Uzuner, O., & Szolovits, P. (2017). *De-identification of Patient Notes with Recurrent Neural Networks*. Journal of the American Medical Informatics Association.
4. Pilán, I., Volodina, E., & Alfter, D. (2020). *Detecting Personal Data in Unstructured Text*. Proceedings of LREC.
5. McCallister, E., Grance, T., & Scarfone, K. (2010). *Guide to Protecting the Confidentiality of Personally Identifiable Information (PII)*. NIST Special Publication 800-122.
6. European Union. (2016). *General Data Protection Regulation (GDPR)*. Regulation (EU) 2016/679.
7. Smith, R. (2007). *An Overview of the Tesseract OCR Engine*. Proceedings of ICDAR.
8. Microsoft. (n.d.). *Microsoft Presidio: Data Protection and De-identification SDK*. GitHub Repository.
9. PyCA Cryptography. (n.d.). *Fernet (Symmetric Encryption)*. Official Documentation.