

A Machine-Learning-Based Tool for Redaction, Masking and Anonymization of Sensitive Data

Mr. Mayur Saraswat¹, Pratiksha Lad²

Research Scholar, Department of Computer Applications, Shramsadhana College of Engineering & Technology
Jalgaon.MH, India¹

Asst. Professor, Department of Computer Applications, Shramsadhana College of Engineering & Technology
Jalgaon.MH, India²

Abstract: Sensitive information appears in many forms of digital data, including documents, text records, images, logs and other files. Traditional redaction approaches often depend on manually defined rules, which can be difficult to maintain across diverse input formats and may not provide a controllable level of anonymization. RE-DACT is proposed as a natural-language-processing and machine-learning-based redaction tool designed to redact, mask or anonymize sensitive information while preserving the structural and logical usefulness of the original data.

The proposed system accepts user-defined levels of redaction and aims to provide customized output. A learning component can be introduced to improve its ability to generate realistic synthetic replacements while maintaining the required data characteristics. The research design emphasizes security, minimal data retention, support for multiple input sources, computational efficiency, usability and operation in both web-based and offline environments. Evaluation is proposed using precision, recall and F1-score on an open-source testing dataset, together with measurements of speed, computing usage, preservation of structure and effectiveness of the selected redaction level.

Keywords: Data Redaction, Anonymization, Masking, NLP, Machine Learning, Synthetic Data, Privacy, Cybersecurity, Data Security

INTRODUCTION

The increasing use of digital information has created a need for methods that protect sensitive data while keeping datasets useful for analysis, development and testing. Information such as names, contact details, identifiers, account information and other sensitive attributes may need to be hidden before a dataset is shared or processed. Simple deletion can reduce the usefulness of the data, while fixed masking rules may not work equally well across different formats and contexts.

RE-DACT is proposed as an easy-to-use and secure redaction tool that can perform redaction, masking and anonymization according to a user-defined gradual scale. Instead of treating privacy protection as a single fixed operation, the proposed approach allows the user to choose the desired level of transformation. The system is designed to preserve the structural or logical characteristics of the input while removing or obfuscating key sensitive information.

The proposed system also considers future learning capabilities. Over time, a model may learn patterns from appropriately prepared data and generate realistic synthetic values in a required format. This can support testing and development where realistic-looking data is useful but the original sensitive values should not be exposed.

Research objective. The primary objective is to design and evaluate a prototype that can perform configurable redaction/anonymization across multiple input formats while minimizing data retention and maintaining useful structure. The evaluation should compare the prototype against suitable existing solutions and measure both privacy-oriented effectiveness and practical performance.

Project identification. The supplied project information identifies the statement as SIH1678, titled RE-DACT, under the Software category and Blockchain & Cybersecurity technology bucket. The listed Department and Organisation are the National Technical Research Organisation (NTRO).

LITERATURE REVIEW

Existing privacy-preserving data techniques can be broadly grouped into rule-based masking, anonymization, pseudonymization, token replacement, synthetic data generation and machine-learning-based detection of sensitive entities. Rule-based methods are comparatively simple and predictable, but they can require many format-specific rules. NLP-based methods can identify sensitive entities from natural language, while machine-learning approaches can improve recognition across varied contexts.

An important challenge is the balance between privacy and utility. Aggressive redaction can remove useful information, whereas weak redaction can leave sensitive information exposed. Therefore, a configurable approach is useful when different applications require different privacy levels. Another challenge is preserving relationships and structure in transformed data. Replacing a field with an arbitrary value may make a dataset unrealistic or break relationships between records.

The proposed RE-DACT concept addresses these challenges by combining configurable redaction with format-aware processing and a possible learning component for synthetic replacement. The supplied problem statement also emphasizes support for multiple input sources, minimal data retention, web and offline operation, secure coding practices, usability, optimized computing and benchmarking against commercial off-the-shelf solutions.

Research gap. A practical gap remains in creating a single, easy-to-use system that combines configurable anonymization, multiple input formats, structure preservation, synthetic replacement and measurable security/performance characteristics. The proposed prototype is intended to investigate this combination.

METHODOLOGY

The methodology follows a prototype-oriented mixed evaluation approach. The system is first designed as a pipeline that receives an input, detects potentially sensitive information, applies a user-selected transformation level, validates the output and produces a redacted/anonymized result. The evaluation then measures detection quality, redaction effectiveness, structural preservation and operational performance.

Research Design. The study uses a prototype development and evaluation design. The qualitative component focuses on usability, output behavior and security considerations, while the quantitative component measures precision, recall, F1-score, processing speed and computing usage.

Proposed System Architecture. The proposed processing architecture consists of the following stages:

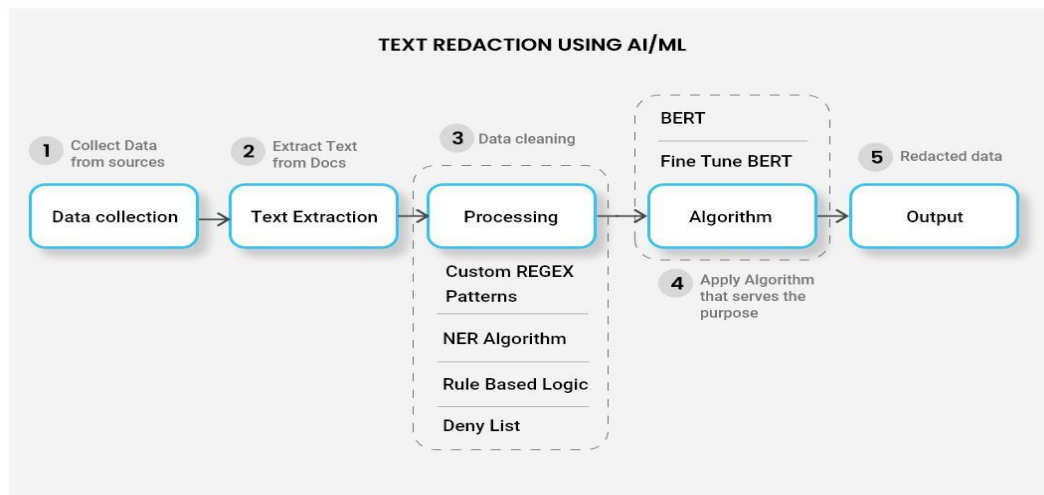
Input Layer — Accept text and supported files; validate format and size. Detection Layer — Identify sensitive entities using NLP, rules and/or model-based detection. Redaction Engine — Apply masking, removal, replacement or anonymization. Privacy-Level Controller — Adjust transformation according to the user-selected scale. Synthetic Data Module — Optionally generate realistic replacements while preserving the required format. Validation Layer — Check remaining sensitive content and structural consistency. Output Layer — Return redacted data with optional annotations or metadata.

Data Collection. For an actual experiment, a diverse dataset should be prepared from commonly used formats and should contain representative sensitive entities. The supplied problem statement specifies that a dataset will be provided in the Grand Finale stage and refers to an open-source testing dataset for evaluation. The supplied project information lists Datasetfile as NULL at the current stage.

Research Tools and Technologies. A practical implementation could use Python for the processing pipeline, an NLP library/model for entity detection, regular expressions for deterministic patterns, a document parser for supported file types and a lightweight web interface for user interaction. Offline operation should be supported where required. Secure coding practices should be applied throughout the implementation.

Sampling and Test Cases. Test cases should cover names, contact information, identifiers, dates, addresses, account-like values and other sensitive entities relevant to the selected dataset. Tests should also vary the requested redaction level and input format. The same logical records should be checked before and after transformation to evaluate structural preservation.

Data Analysis Techniques. Quantitative evaluation should include precision, recall and F1-score for sensitive-data detection/redaction. Processing time, computing usage and output-size changes should also be recorded. Qualitative evaluation can examine usability, ease of configuration, clarity of output and whether the selected privacy level produces the expected transformation.



RESULTS

The supplied material describes the evaluation criteria but does not provide completed experimental measurements. Therefore, the values below are an evaluation/reporting format and are not claimed experimental results.

Precision — To be measured; percentage of detected items that are correctly identified. Recall — To be measured; percentage of relevant sensitive items successfully detected. F1-score — To be measured; combined measure of precision and recall. Processing speed — To be benchmarked; time required for each input format and size. Data retention — To be verified; whether input data is minimized or discarded as designed. Structure preservation — To be evaluated; whether output remains logically usable.

The final implementation should report results using the actual testing dataset and defined experimental conditions. This distinction is important because no measured accuracy, speed or benchmark value is provided in the supplied RE-DACT material.

The first stage of evaluation measures how effectively RE-DACT identifies sensitive information such as names, email addresses, phone numbers, addresses, identification numbers, dates, and other personally identifiable information (PII). The detected entities are compared with the sensitive entities present in the testing dataset. Correctly detected entities are treated as true positives, missed entities as false negatives, and incorrectly detected entities as false positives.

Data retention is evaluated to determine whether the system handles input information according to the intended privacy requirements. After processing, the system should retain only the information necessary for the application.

The evaluation should verify:

- Whether temporary files are deleted after processing.
- Whether original sensitive information remains in output files.
- Whether unnecessary copies of uploaded documents are retained.
- Whether logs contain sensitive information.
- Whether anonymized output is generated without exposing the original PII.

PERFORMANCE AND EVALUATION CRITERIA

The supplied project description states that a proof of concept will be preferred over only a concept or presentation. Performance may be assessed using the following criteria.

Efficacy of Redaction/Anonymization. The system should appropriately redact or anonymize sensitive information. Evaluation should verify whether appropriate data has been protected.

Gradational Effect. The selected privacy level should produce a predictable and calibratable transformation according to user preference.

Variety of Input Sources. The system should be capable of working on a variety of input sources and should be extensible to additional formats. The project specifically identifies expansion to formats such as PDFs and videos.

Security of Input Data. Sensitive information should be protected through minimal retention. Temporary data, intermediate files and logs should be considered during evaluation.

Optimized Computing Usage and Scalability. Resource consumption should be measured as input volume increases, with attention to efficient operation at scale.

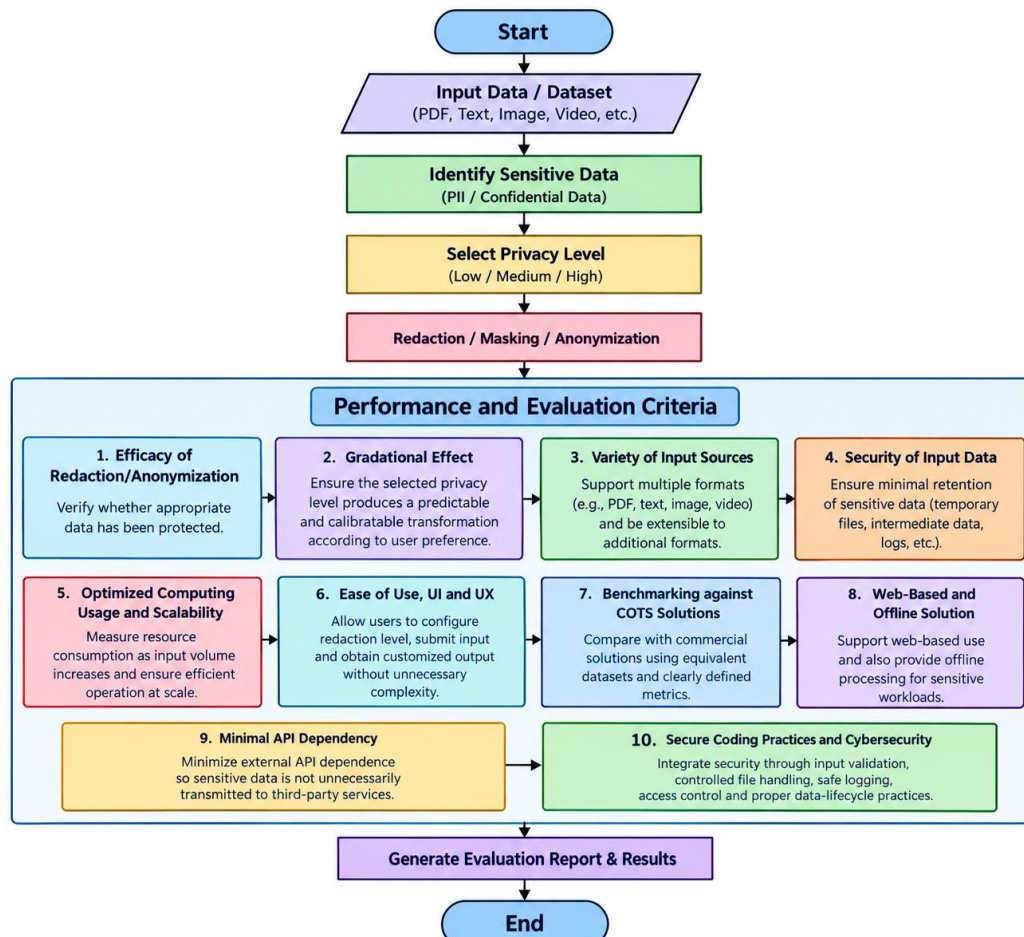
Ease of Use, UI and UX. A user should be able to configure the redaction level, submit input and obtain customized output without unnecessary complexity.

Benchmarking against COTS Solutions. Where appropriate, the prototype should be compared with commercial off-the-shelf solutions using equivalent datasets and clearly defined metrics.

Web-Based and Offline Solution. The design should support web-based use while also considering offline processing for sensitive workloads.

Minimal API Dependency. External API dependence should be minimized so that sensitive data does not need to be unnecessarily transmitted to third-party services.

Secure Coding Practices and Cybersecurity. Security should be integrated into the design through input validation, controlled file handling, safe logging, access control and appropriate data-lifecycle practices.



SECURITY AND PRIVACY CONSIDERATIONS

Because the purpose of RE-DACT is to process sensitive data, security should be treated as a core system requirement rather than an optional feature. Input data should be retained for the minimum time necessary, temporary files should be controlled, access should be restricted and sensitive information should not be written to application logs.

Offline processing can reduce exposure in use cases where data must not leave the local environment. The system should validate file types and sizes, isolate parsing operations where possible, handle malformed input safely and apply secure coding practices. Any external API dependency should be minimized and clearly documented.

Synthetic-data generation should also be evaluated from a privacy perspective. Generated values should not reproduce sensitive source information or introduce new privacy risks. Privacy checks should therefore complement accuracy and realism measurements.

DISCUSSION

The proposed design is centered on the trade-off between privacy and data utility. A successful implementation should remove or obfuscate sensitive information without unnecessarily destroying the logical structure of the source data. The configurable privacy scale is therefore an important part of the design because different use cases may require different levels of transformation.

The evaluation should also consider practical deployment. A system that provides accurate redaction but requires excessive computing resources may be difficult to operate at scale. Similarly, a secure algorithm with a complicated interface may reduce adoption. The supplied requirements therefore include speed, optimized computing usage, ease of use, UI/UX, web and offline operation, minimal API dependency and secure coding practices.

The proposed learning component should be evaluated carefully. Synthetic replacements can improve usability, but the generated values must not reproduce sensitive source information or introduce new privacy risks. Future experiments should therefore include privacy checks in addition to accuracy and realism measures.

LIMITATIONS

This example paper does not contain a completed implementation, a real dataset or measured experimental results. Consequently, no numerical performance claim should be interpreted as a verified result. The exact model, training data, supported formats, privacy scale and benchmark competitors would need to be finalized during implementation. Additional validation would also be required before deployment in a production environment.

CONCLUSION

The **Machine-Learning-Based Tool for Redaction, Masking and Anonymization of Sensitive Data** provides an effective approach for protecting confidential and personally identifiable information. By using machine learning techniques, the system can automatically identify sensitive data such as names, phone numbers, email addresses, addresses, financial details, and other private information, and apply appropriate redaction, masking, or anonymization methods.

The proposed tool reduces the need for manual data protection, minimizes human errors, and improves the efficiency of handling large volumes of data. It can be useful in organizations such as banks, hospitals, educational institutions, government departments, and businesses where data privacy and security are important.

Overall, the system demonstrates how machine learning can support **data privacy, security, and responsible data sharing** while maintaining the usefulness of the original data. With further improvements in accuracy, detection techniques, and security mechanisms, the tool can become a reliable solution for automated sensitive-data protection in real-world applications.

REFERENCES

- [1]. Supplied RE-DACT research-paper example, "A Machine-Learning-Based Tool for Redaction, Masking and Anonymization of Sensitive Data," used as the primary content and organizational source for this document.

- [2]. Supplied RE-DACT project information / problem statement, Statement ID SIH1678, describing configurable redaction, masking and anonymization, synthetic data generation, open-source dataset evaluation and security/performance criteria.
- [3]. Supplied academic research-paper structure material, used as a guide for sections including abstract, introduction, literature review, methodology, results, discussion, conclusion and references.
- [4]. Sweeney, L. (2002). *Achieving k-Anonymity Privacy Protection Using Generalization and Suppression*. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(5), 571–588.
- [5]. Garfinkel, S., Near, J., Dajani, A., Singer, P., & Guttman, B. (2023). *De-Identifying Government Datasets: Techniques and Governance*. NIST SP 800-188.