Evolutionary Algorithm Approach for Detecting Personal and Protected Health Information in Unstructured Data

Syed Momin Naqvi, Rehmat Gul
Institute of Business Administration, Karachi

April 18, 2025

Abstract

This research proposes a novel approach using genetic algorithms to detect Personal Identifiable Information (PII) and Protected Health Information (PHI) in diverse unstructured text sources. Unlike traditional methods that rely on predefined rules or supervised learning requiring extensive labeled data, our approach evolves detection patterns through natural selection principles, potentially offering greater adaptability to new data variations and previously unseen patterns.

1 Introduction

1.1 Problem Statement

Despite advances in natural language processing and machine learning, existing PII/PHI detection systems face significant challenges:

- Limited adaptability to new or variant expressions of sensitive information
- Reliance on extensive human-annotated training data
- Rigidity in rule-based systems that require manual updates

These limitations create a need for more flexible approaches that can evolve with changing data patterns without extensive retraining or manual rule engineering.

1.2 Research Objectives

This research aims to develop and evaluate a genetic algorithm approach to detect PII and PHI in unstructured data by:

- Designing an appropriate chromosome representation for PII/PHI detection patterns
- Developing specialized fitness functions and genetic operators
- Evaluating the approach against traditional methods across multiple domains

2 Literature Review

2.1 Traditional PII/PHI Detection Methods

Current approaches include rule-based systems, machine learning models, and hybrid methods. Rule-based systems achieve high precision for well-defined PHI categories but struggle with context-dependent information. Machine learning approaches offer improved flexibility but require substantial labeled data. Hybrid approaches combining rules and machine learning show promise but often have significant computational overhead.

2.2 Evolutionary Approaches in Text Analysis

While genetic algorithms have not been extensively applied to PII/PHI detection specifically, they have shown promise in related domains:

- Pattern discovery in text data
- Feature selection for text classification
- Adaptive text processing systems that transfer between domains

2.3 Research Gap

There remains a gap in applying evolutionary approaches specifically to PII/PHI detection. Potential advantages include reduced reliance on labeled data, improved adaptability to new patterns, and continuous evolution without explicit reprogramming.

3 Methodology

3.1 Chromosome Representation

We propose a multi-gene chromosome structure where each gene represents a detection pattern for a specific type of PII/PHI, containing:

- Pattern component (regex or feature-based)
- Context parameters

- PII/PHI type identifier
- Confidence parameters

3.2 Genetic Operators

Specialized operators will include:

- **Mutation:** Pattern expansion/restriction, context modification
- **Crossover:** Gene exchange, pattern merging, ensemble creation

3.3 Fitness Function

The fitness function will balance multiple objectives:

$$\text{Fitness} = \alpha \cdot \text{Precision} + \beta \cdot \text{Recall} + \gamma \cdot \text{Novelty} - \delta \cdot \text{Complexity} \tag{1}$$

4 Experimental Design

4.1 Datasets and Evaluation

We will evaluate using multiple datasets (clinical notes, emails, legal documents) against baselines including rule-based systems, traditional ML, and deep learning approaches. Performance will be measured using precision, recall, F1 score, adaptation speed, and transfer performance.

4.2 Experimental Scenarios

We will test the system under various scenarios:

- Standard detection with comprehensive training data
- Few-shot learning with limited examples
- Cross-domain adaptation
- Novel pattern detection

5 Expected Outcomes and Impact

5.1 Technical Contributions

- Novel chromosome representation for PII/PHI detection patterns
- Specialized genetic operators for text pattern evolution
- Empirical evaluation of evolutionary approaches versus traditional methods

5.2 Potential Applications

Applications include privacy-preserving document processing, compliance monitoring, automated data sanitization, and systems that adapt to evolving privacy requirements.

5.3 Limitations and Mitigations

Anticipated challenges include computational complexity, initial knowledge requirements, and explainability issues. We will address these through optimizations, initial population seeding, and techniques to favor simpler patterns.

6 Conclusion

This research proposes a novel evolutionary approach to PII/PHI detection in unstructured data. By applying genetic algorithms to evolve detection patterns, we aim to develop a system that can adapt to new data variations with minimal human intervention, potentially improving both effectiveness and efficiency across diverse domains.

References

References

[1] Neumann, A., et al. (2019). "Implementing rule-based de-identification methods for protected health information in clinical text." BMC Medical Informatics and Decision Making, 19(1), 1-11.

[2] Chen, T., & Thompson, P. (2020). "Transformer-based PHI detection in clinical notes with limited training data." Proceedings of the Conference on Health, Inference, and Learning, 87-96.

[3] Zhao, L., et al. (2021). "Hybrid rule-based and neural approaches for protected health information recognition." Journal of Biomedical Informatics, 118, 103791.

[4] Fernandez, J. G., & García, M. (2018). "Evolutionary pattern discovery in system logs." IEEE Transactions on Knowledge and Data Engineering, 30(3), 559-572.

[5] Rodriguez, T., et al. (2019). "Evolutionary adaptation of text processing rules for domain-specific information extraction." Applied Soft Computing, 85, 105765.