# Anonymization of Personal Data in Unstructured Documents Using Natural Language Processing and Machine Learning: A Systematic Literature Review

Marisa Moreno

## Planning

Identificar, analisar e sintetizar estudos publicados na área de segurança pública que aplicaram aprendizado de máquina e processamento de linguagem natural para a identificação e anonimização de dados pessoais — incluindo nomes, e-mails, números de telefone, números de documentos e dados bancários — em documentos textuais não estruturados, especialmente boletins de ocorrência policial. A revisão avaliará a eficácia dessas técnicas por meio de métricas como precisão, acurácia, recall e F1-score, bem como seu impacto na preservação da integridade e utilidade dos textos originais, considerando a conformidade com legislações de proteção de dados, como a Lei Geral de Proteção de Dados (LGPD).

### PICOC

* **Population:** Unstructured textual documents in the field of public security, especially police reports.
* **Intervention:** Machine learning and natural language processing techniques for identifying and anonymizing personal data.
* **Comparison:** Traditional and hybrid models for recognizing and anonymizing personal data (e.g., BERTimbau, spaCy, rule-based systems).
* **Outcome:** EEffectiveness of anonymization measured by statistical metrics; preservation of the integrity and utility of the original texts; compliance with data protection regulations such as the GDPR and LGPD.
* **Context:** Anonimização de dados pessoais para garantir conformidade com a LGPD em contextos de segurança pública.

### Research Questions

1. Q1 – Quais algoritmos de aprendizado de máquina têm sido aplicados para identificar e anonimizar diferentes tipos de dados pessoais (nomes, e-mails, telefones, documentos, dados bancários) em documentos textuais não estruturados?
2. Q2 – Quais técnicas de Processamento de Linguagem Natural são empregadas para apoiar a identificação e anonimização desses dados pessoais em textos não estruturados?
3. Q3 – Quais abordagens computacionais são utilizadas para lidar com desafios como ambiguidades semânticas, variações linguísticas e dependência de contexto na anonimização automática de dados pessoais?
4. Q4 – Quais métodos baseados em análise gramatical e contextual são usados para melhorar a precisão e o recall na anonimização de dados pessoais?
5. Q5 – Quais métricas e métodos de avaliação são utilizados para mensurar a eficácia dos modelos de anonimização, considerando tanto a performance técnica quanto a preservação da integridade e utilidade dos textos originais?
6. Q6 – De que maneira essas abordagens são validadas e aplicadas em contextos reais da segurança pública, especialmente em documentos como boletins de ocorrência policial?

### Keywords and Synonyms

|  |  |
| --- | --- |
| Keyword | Synonyms |
| Data privacy | data utility, information preservation |
| Data protection laws | GDPR, LGPD, privacy regulations |
| Machine learning | artificial intelligence, machine learning algorithms |
| Natural language processing | text mining |
| Personal data anonymization | data masking, de-identification, identity masking |
| Personal data recognition models | BERTimbau, hybrid models, rule-based systems, spaCy, traditional models |
| Police reports | crime reports, law enforcement text, police narratives |
| Unstructured documents | free-text documents, legal texts, police reports, textual data, unstructured texts |

### Search String

("police reports" OR "crime reports" OR "law enforcement text" OR "police narratives" OR "unstructured documents" OR "free-text documents" OR "legal texts" OR "textual data" OR "unstructured texts") AND ("machine learning" OR "artificial intelligence" OR "machine learning algorithms" OR "natural language processing" OR "text mining" OR "personal data anonymization" OR "data masking" OR "de-identification" OR "identity masking") AND ("personal data recognition models" OR "BERTimbau" OR "hybrid models" OR "rule-based systems" OR "spaCy" OR "traditional models") AND ("data privacy" OR "data utility" OR "information preservation" OR "data protection laws" OR "GDPR" OR "LGPD" OR "privacy regulations")

### Sources

* ACM Digital Library (http://portal.acm.org)
* IEEE Digital Library (http://ieeexplore.ieee.org)
* Scopus (http://www.scopus.com)
* Springer Link (http://link.springer.com)

### Selection Criteria

**Inclusion Criteria:**

* IC1: O estudo aplica técnicas de aprendizado de máquina e/ou processamento de linguagem natural (PLN) para anonimização de dados pessoais.
* IC2: O estudo considera características contextuais e/ou gramaticais para melhorar a identificação de nomes pessoais.
* IC3: O estudo analisa textos não estruturados, especialmente boletins de ocorrência ou documentos com alta variabilidade linguística.
* IC4: O estudo aborda desafios típicos de textos não estruturados, como ambiguidade semântica, variações regionais ou erros de digitação.
* IC5: O estudo apresenta resultados experimentais com métricas de avaliação, como precisão, revocação (recall), F1-score, e discute o impacto sobre a integridade do texto.
* IC6: O estudo menciona leis de proteção de dados (ex.: LGPD, GDPR) ou discute aspectos de privacidade e segurança pública relacionados à anonimização.

**Exclusion Criteria:**

* EC1. O estudo está escrito em idioma diferente do inglês.
* EC10. O estudo não aborda estratégias para reduzir ambiguidades contextuais ou gramaticais, importantes na identificação precisa de dados pessoais em linguagem natural.
* EC2. O estudo foi publicado antes de 2020, não refletindo o contexto atual da LGPD
* EC3. O estudo não utiliza aprendizado de máquina nem processamento de linguagem natural, baseando-se apenas em técnicas estatísticas tradicionais ou regras manuais sem inteligência computacional.
* EC4. O estudo não tem como foco a anonimização de dados pessoais, como nomes, e-mails, telefones, documentos ou dados bancários.
* EC5. O estudo não aborda textos não estruturados, ou utiliza exclusivamente dados estruturados.
* EC6. O estudo é uma revisão teórica ou estudo conceitual, sem realização de experimentos ou validação prática.
* EC7. O estudo não apresenta métricas quantitativas de avaliação u não discute impacto sobre a integridade/uso dos textos após anonimização.
* EC8. O estudo não utiliza ou não considera dados oriundos da área da segurança pública, como boletins de ocorrência ou documentos similares.
* EC9. O estudo não menciona ou não discute conformidade com legislações de proteção de dados, como a LGPD, GDPR ou princípios relacionados à privacidade e segurança da informação.

### Quality Assessment Checklist

**Questions:**

**Answers:**

### Data Extraction Form

## Conducting

### Digital Libraries Search Strings

**ACM Digital Library:**

[[All: "police reports"] OR [All: "crime reports"] OR [All: "law enforcement"] OR [All: "police narratives"]] AND [All: and] AND [[All: "machine learning"] OR [All: "artificial intelligence"] OR [All: "natural language processing"] OR [All: "text mining"] OR [All: "personal data anonymization"]] AND [All: and] AND [[All: "hybrid models"] OR [All: "rule-based systems"] OR [All: "personal data recognition models"] OR [All: spacy] OR [All: bertimbau]] AND [All: and] AND [[All: "data privacy"] OR [All: gdpr] OR [All: lgpd] OR [All: "data protection laws"] OR [All: "privacy regulations"]] AND [E-Publication Date: Past 5 years]

**IEEE Digital Library:**

("police reports" OR "crime reports" OR "law enforcement" OR "unstructured text" OR "free-text" OR "legal text")  
  
AND  
  
("machine learning" OR "natural language processing" OR "text mining" OR "personal data anonymization" OR "de-identification")

**Scopus:**

TITLE-ABS-KEY ( ( "police reports" OR "crime narratives" OR "unstructured documents" OR "free-text" OR "legal texts" ) AND ( "machine learning" OR "artificial intelligence" OR "natural language processing" OR "text mining" ) AND ( "personal data anonymization" OR "data masking" OR "de-identification" OR "data privacy" OR "GDPR" OR "LGPD" ) )

**Springer Link:**

(ALL:("police reports" OR "crime reports" OR "law enforcement text" OR "police narratives" OR "unstructured documents" OR "free-text documents" OR "legal texts" OR "textual data" OR "unstructured texts")) AND (ALL:("machine learning" OR "artificial intelligence" OR "machine learning algorithms" OR "natural language processing" OR "text mining" OR "personal data anonymization" OR "data masking" OR "de-identification" OR "identity masking")) AND (ALL:("personal data recognition models" OR "BERTimbau" OR "hybrid models" OR "rule-based systems" OR "spaCy" OR "traditional models")) AND (ALL:("data privacy" OR "data utility" OR "information preservation" OR "data protection laws" OR "GDPR" OR "LGPD" OR "privacy regulations"))

### Imported Studies

* **ACM Digital Library:** 6
* **IEEE Digital Library:** 1543
* **Scopus:** 34
* **Springer Link:** 0