
# Anonimização de Incidentes de Segurança com Reidentificação Controlada

**Carolina Tompsen Bandel[2], João Pedro Ramires Esteves[3], Kalian Pereira Guerra[1], Leandro M. Bertholdo[1] Diego Kreutz[3], Rodrigo S. Miani[4],**

[1] Universidade Federal do Rio Grande do Sul (UFRGS)
[2]Centro Universitário SENAC (SENAC EAD)
[3]Universidade Federal do Pampa (UNIPAMPA)
[4]Universidade Federal de Uberlândia (UFU)

{leandro.bertholdo, kalian.guerra}@ufrgs.br,
carolina.tbandel@senacsp.edu.br, joao.esteves@ufu.br,
diegokreutz@unipampa.edu.br,
miani@ufu.br

***Resumo.*** *Este trabalho apresenta o AnonLFI, um framework híbrido para anonimizar relatos de incidentes de segurança escritos em linguagem natural. A proposta busca viabilizar o uso desses dados por Large Language Models (LLMs), ao mesmo tempo em que protege informações sensíveis da própria IA. O framework combina pseudoanonimização determinística, reconhecimento de entidades nomeadas (NER) e expressões regulares para tratar dados não estruturados, preservando o contexto original e permitindo reidentificação controlada quando necessário. O trabalho define requisitos específicos para anonimização reutilizável no contexto de CSIRTs, avalia ferramentas existentes e demonstra, com base na análise de 763 incidentes reais, que o framework alcança 97,38% de eficácia sem gerar falsos positivos. Os resultados demonstram a eficácia da ferramenta em anonimizar dados de incidentes de segurança reais, preservando o contexto e a utilidade das informações para análise.*

## 1. Introdução

O aumento da complexidade das ameaças cibernéticas tem tornado o tratamento de incidentes de segurança uma atividade que ultrapassa a mera resposta técnica, exigindo colaboração entre múltiplos atores. Em cenários como a detecção de *botnets*, que envolvem diversos domínios e instituições, a ausência de compartilhamento estruturado de informações entre *Computer Security Incident Response Teams* (CSIRTs) compromete a identificação coordenada de ataques. Ampliar essa colaboração pode fortalecer a capacidade coletiva de resposta e contribuir para uma segurança cibernética mais eficaz. Entretanto, proteger a privacidade de dados sensíveis presentes na descrição dos incidentes de segurança é considerado o principal desafio dessa colaboração, visando evitar o vazamento de dados como IPs, endereços MAC, e-mails, nomes de domínio, identificadores de usuários e senhas.

Neste contexto, a anonimização de dados emerge como uma solução indispensável para ampliar a colaboração entre CSIRTs e viabilizar o processamento automatizado de incidentes em conformidade com normas e legislações vigentes, como

a Lei Geral de Proteção de Dados (LGPD). No entanto, as técnicas convencionais [Majeed and Lee 2020, Murthy et al. 2019], como a anonimização irreversível, frequentemente se mostram insuficientes para lidar com a natureza não estruturada dos relatos de segurança, que incorporam linguagem técnica, acrônimos específicos e metadados confidenciais. Adicionalmente, a aplicação de técnicas recentes [Yang et al. 2024], como as baseadas em técnicas de Inteligência Artificial (IA), para a análise automatizada de incidentes através de *Large Language Models* (LLMs) exige dados de treinamento ricos em contexto, mas desprovidos de informações identificáveis.

O estado da arte em técnicas de anonimização para cibersegurança revela duas abordagens primárias: a anonimização irreversível, que remove permanentemente identificadores como IPs e endereços MAC [Aleroud et al. 2021], e a pseudonimização, que emprega técnicas de *hash* ou criptografia para permitir a reidentificação controlada [Varanda et al. 2021]. A pseudonimização demonstra relevância particular para análise de segurança em *Managed Security Service Providers* (MSSPs), onde a rastreabilidade dos dados e a estrutura de domínio são necessárias para investigar ataques distribuídos [Zhang et al. 2006]. Desafios significativos persistem na heterogeneidade dos formatos de *logs* (estruturados e não estruturados) e na necessidade de preservar características realistas dos dados pós-anonimização [Koukis et al. 2006]. Consequentemente, abordagens híbridas que combinam diferentes técnicas, aliadas ao uso de Inteligência Artificial (IA), apresentam-se como alternativas promissoras para lidar com casos complexos—como a análise de dados não estruturados presentes em relatos de incidentes de cibersegurança —visando sua posterior ingestão em sistemas de *cyber threat intelligence*.

Diante desse panorama, este trabalho propõe uma abordagem sistemática e prática de anonimização de incidentes de segurança, concretizada através da implementação do *framework* **AnonLFI**. Esta solução busca harmonizar a proteção de dados pessoais com a manutenção do valor analítico dos incidentes reportados. Foi adotada uma abordagem híbrida utilizando técnicas avançadas como pseudoanonimização, processamento de linguagem natural e reconhecimento de padrões via expressões regulares (*regex*), em consonância com o padrão internacional IODEF (*Incident Object Description Exchange Format*)[1] para assegurar a interoperabilidade entre equipes de resposta a incidentes.

As contribuições do AnonLFI se manifestam em três dimensões cruciais. Primeiramente, o desenvolvimento do *framework* representa uma inovação tangível no domínio da cibersegurança, oferecendo um processo automatizado local para a transformação segura de dados sensíveis. Em segundo lugar, a aplicação prática em ambientes reais demonstra a viabilidade da solução, utilizando dados reais de incidentes, enquanto salvaguarda a privacidade das entidades envolvidas. Por fim, o trabalho estabelece diretrizes claras para a adoção institucional, facilitando a transição de organizações para um paradigma de compartilhamento seguro de informações sobre ameaças cibernéticas.

## 2. Métodos de anonimização aplicados em cibersegurança

Anonimização é o processo de remoção de identificadores, preservando a utilidade analítica dos dados, o que exige atenção especial a atributos como endereços IP, dados pessoais, e nomes de *hosts* e domínios, a fim de evitar reidentificação [Fisk et al. 2015, Senavirathne and Torra 2020, Aleroud et al. 2021, Yang et al. 2024]. A preservação da

[1] https://www.ietf.org/rfc/rfc5070.txt

estrutura de domínio desses dados é fundamental para mitigar ataques por sondagem, especialmente em cenários de terceirização da análise de segurança [Zhang et al. 2006]. Uma alternativa para lidar com esse desafio e manter correlações relevantes entre os dados é a pseudoanonimização [Varanda et al. 2021], que permite a reidentificação controlada por meio de mecanismos como funções de hash criptográfica, criptografia e tabelas de correspondência (*rainbow tables*), sendo particularmente útil em ambientes que exigem rastreabilidade.

Diversas abordagens de anonimização foram propostas para proteger dados sensíveis sem comprometer sua utilidade. No caso de incidentes de segurança, a escolha da técnica deve considerar a complexidade e a variedade dos dados envolvidos, como *logs* e descrições técnicas. Entre as principais técnicas de anonimização podemos relacionar [Imperva 2025, Tempest 2021, Majeed and Lee 2021]:

1. **Generalização**: excluir parte dos dados para torná-los menos específicos, impossibilitando uma identificação direta. Exemplo: substituir "`Rua das Flores, Umuarama, 42`" por "`Rua das Flores`".
2. **Agregação**: agrupar atributos em categorias. Exemplo: converter "`25 anos`" para "`20--30 anos`".
3. **Permutação**: embaralhar valores de atributos sensíveis entre registros semelhantes para quebrar relações diretas.
4. **Mascaramento/Supressão**: ocultar parte das informações com caracteres como '`*`' ou '`x`'. Exemplo: "`987.***.***-**`".
5. **Pseudoanonimização**: substituir dados reais por identificadores consistentes e reversíveis via hash ou chaves. Exemplo: "`João Silva`" por "`hash-pessoa-001`".
6. **Perturbação**: adicionar ruído para preservar padrões estatísticos. Exemplo: alterar "`R$ 1.000,00`" para "`R$ 1.012,34`".
7. **Anatomização**: separar atributos identificadores dos sensíveis em duas tabelas distintas, quebrando a associação direta entre eles.

Com o passar do tempo, estas técnicas evoluíram impulsionadas por desafios como lidar com grandes volumes de dados, automação e conformidade regulatória de segmentos específicos, como o caso dos provedores de acesso à Internet (ISPs) [Portillo-Dominguez and Ayala-Rivera ]. Em contextos de tráfego em larga escala, foram investigadas estratégias voltadas à adequação a normas como a *General Data Protection Regulation* (GDPR), combinando anonimização, escalabilidade e mecanismos de consentimento informado [Fejrskov et al. 2020]. Outros equilibram privacidade e realismo dos dados, com foco na manipulação de *logs* e rastros de rede, como *network flows* [Rasic 2020, Koukis et al. 2006].

Além destas técnicas e dos formatos de dados tratados, é relevante considerar a finalidade para a qual cada solução foi concebida. Parte dos trabalhos na área tenta resolver problemas como a terceirização da análise de *logs*; caso dos provedores de serviços gerenciados (MSSPs) [Zhang et al. 2006, Rasic 2020]. Uma abordagem comum aqui, são as técnicas de minimização e supressão de dados [Portillo-Dominguez and Ayala-Rivera ]. Entretanto, nenhuma dessas propostas trata da anonimização de textos desestruturados para aplicação em modelos de linguagem.

Ferramentas como ARX [Prasser et al. 2017], sdcMicro [Templ et al. 2015] e Amnesia [Haber et al. 2022] oferecem suporte à anonimização estruturada, mas não

atendem ao tipo de dado tratado neste trabalho, que envolve descrições textuais não padronizadas. Soluções como CryptoPAn [Xu et al. 2001], FLAIM [Slagell et al. 2006] e cryptopANT [ANT Lab 2018] conseguem manipular também dados não estruturados, mas são restritos a padrões sintáticos reconhecíveis, não operando eficientemente sobre textos em linguagem natural. Por outro lado, ferramentas de Processamento de Linguagem Natural (NLP, *Natural Language Processing*), como spaCy [AI 2025], Presidio [Microsoft 2025] e modelos disponibilizados pela Hugging Face [Wolf et al. 2019], oferecem suporte à identificação de entidades sensíveis via reconhecimento de entidades nomeadas (NER - *Named Entity Recognition*), embora nem sempre atendam aos requisitos de anonimização reutilizável ou permitam a execução local.

Apesar dos avanços, observa-se uma lacuna de propostas voltadas à anonimização de incidentes de segurança descritos em linguagem natural, especialmente em língua Portuguesa e com foco na integração com modelos de linguagem (LLMs). As LLMs são suscetíveis a vazamentos de dados por meio de ataques ativos e passivos, mesmo quando técnicas como aprendizado federado e privacidade diferencial são empregadas [Yan et al. 2024]. Esses modelos podem memorizar e revelar informações pessoais sensíveis, comprometendo a privacidade dos usuários. As estratégias para mitigar esses riscos envolvem intervenções em diferentes fases, como higienização dos dados no pré-treinamento, aplicação de *unlearning* seletivo no *fine-tuning* e uso de métodos criptográficos durante a inferência. No entanto, tais medidas introduzem novos desafios relacionados à transparência dos modelos e à escalabilidade das soluções.

Adicionalmente, estudos recentes [Gunay et al. 2024, Yan et al. 2024] reforçam os riscos de vazamentos de dados pelas LLMs, tanto por memorizar informações sensíveis quanto por permitir sua extração via ataques ativos ou passivos. Tais riscos evidenciam a importância de mecanismos de pré-processamento seguros e reutilizáveis, capazes de anonimizar dados antes que sejam utilizados em modelos. Embora o presente trabalho não utilize LLMs diretamente, a proposta do *framework* visa preparar dados sensíveis para possível uso futuro com esses modelos, atuando como uma etapa intermediária de anonimização.

| **Referência** | **Formato de Dados** | **Técnica** | **Preparação para uso em IA/LLMs** |
|---|---|---|---|
| Zhang et al. | Logs de rede | Ofuscação | ✗ |
| Varanda et al. | Logs e eventos | Pseudoanonimização | ✗ |
| Fejrskov et al. | NetFlow/DNS | Generalização | ✗ |
| Portillo-Dominguez | Logs | Minimização/Anon. | ✗ |
| Rasic | Logs | Pseudoanonimização | ✗ |
| Koukis et al. | Tráfego de rede | Prefix-preserving | ✗ |
| AnonLFI | Texto não estruturado | Hash + NER + RegEx | ✓ |

**Tabela 1. Métodos utilizados para anonimização de dados em ISPs e MSSPs**

O AnonLFI é construído sobre um modelo de linguagem natural do tipo *transformer* (especificamente o *xlm-roberta-base-ner-hrl*) que, embora seja uma rede neural profunda, não se enquadra como uma LLM. Este trabalho propõe preencher essa lacuna por meio de um *framework* híbrido baseado em reconhecimento de entidades nomeadas (NER), expressões regulares (RegEx - *Regular Expression*) e hash determinístico, visando

à anonimização reutilizável de incidentes de segurança (ex.: na forma de *tickets*) e relatórios textuais sensíveis. A Tabela 1 apresenta uma síntese dos métodos revisados, o formato de dados e a técnica de anonimização utilizada.

## 3. Metodologia

A metodologia deste trabalho foi conduzida em três etapas: *(i)* levantamento de requisitos específicos para anonimização de dados sensíveis em incidentes de segurança; *(ii)* avaliação comparativa de ferramentas de anonimização existentes com base nos requisitos definidos; *(iii)* desenvolvimento de um framework híbrido capaz de anonimizar textos não estruturados com preservação de contexto e rastreabilidade; e *(iv)* Avaliação de desempenho baseada em critérios qualitativos e quantitativos.

**Levantamento de Requisitos:** A definição dos requisitos foi guiada por aspectos legais da LGPD, operacionais (interoperabilidade e rastreabilidade entre CSIRTs) e técnicos, incluindo o suporte a dados não estruturados e a utilização dos dados em *prompts* de LLMs. Esses requisitos estão detalhados na Seção 4, e guiaram as decisões de projeto e avaliação ao longo do trabalho.

| Ferramenta | Tipo de Dado | RegExp | NLP | Reversível | T. Livre |
|---|---|---|---|---|---|
| CryptoPAn | IPs de rede estruturados | ✗ | ✗ | △ Parcial | ✗ |
| FLAIM | Logs estruturados | ✓ | ✗ | ✓ | ✓ |
| ip2anonip | Logs contendo IPs | ✗ | ✗ | ✓ | ✗ |
| cryptoANT | IPs em texto livre | ✗ | ✗ | ✓ | △ Parcial |
| LLM-AIx | Prontuários médicos | ✗ | ✓ | ✗ | ✓ |
| spaCy | Textos desestruturados | ✓ | ✓ | ✓ | ✓ |
| Presidio | Textos estruturados e livres | ✓ | ✓ | ✓ | ✓ |
| Hugging Face | Textos desestruturados | ✓ | ✓ | ✓ | ✓ |

**Tabela 2. Ferramentas de Anonimização Avaliadas. Selecionadas em destaque.**

**Avaliação de Ferramentas Existentes:** Foram avaliadas ferramentas tradicionais e modernas de anonimização baseadas em expressões regulares, pseudoanonimização e reconhecimento de entidades nomeadas (NER). Entre elas estão o CryptoPAn [Xu et al. 2001], voltado à anonimização de endereços IP com preservação de prefixos; o FLAIM [Slagell et al. 2006], modular e voltado a logs estruturados; o ip2anonip [Plonka 2003] e o cryptoANT [ANT Lab 2018], voltados à anonimização reversível de IPs em texto; além de soluções baseadas em linguagem natural, como LLM-AIx [Wiest et al. 2024], spaCy [AI 2025], Presidio [Microsoft 2025] e Hugging Face [Wolf et al. 2019]. Cada ferramenta foi testada com mais de 700 relatos reais de incidentes de segurança, oriundos de diferentes CSIRTs acadêmicos, com o objetivo de verificar sua eficácia prática. As limitações observadas auxiliaram no refinamento dos requisitos inicialmente definidos. A Tabela 2 resume as ferramentas avaliadas, destacando aquelas compatíveis com os requisitos estabelecidos, como o suporte a textos não estruturados, detalhados na Seção 4.

**Desenvolvimento do Framework Híbrido:** A partir dos requisitos identificados, foi desenvolvido o `AnonLFI`, um *framework* híbrido para anonimização de incidentes de

segurança. O *framework* integra técnicas de pseudoanonimização determinística, RegEx e NLP, utilizando modelos locais de NER para preservar o contexto e anonimizar entidades sensíveis com rastreabilidade (vide Tabela 2). Detalhes da arquitetura do sistema, algoritmos e tecnologias utilizadas são apresentados na Seção 5.

**Avaliação da Ferramenta:** A ferramenta deve ser avaliada baseado em uma série de critérios qualitativos e quantitativos derivados dos requisitos previamente estabelecidos, preferencialmente avaliados por um analista de segurança. A avaliação quantitativa e qualitativa do desempenho da ferramenta é abordada na Seção 6.

## 4. Requisitos de Anonimização para Incidentes de Segurança em CSIRTs

A anonimização de incidentes de segurança tratados por CSIRTs deve contemplar tanto aspectos técnicos quanto organizacionais, garantindo rastreabilidade, interoperabilidade entre times e preservação de utilidade analítica. A seguir são elencados os requisitos específicos para a anonimização de *tickets* e relatórios de incidentes no contexto da cooperação entre CSIRTs.

**(R1) Pseudoanonimização determinística e reutilizável.** Para permitir a correlação de eventos em diferentes tickets e ao longo do tempo, os identificadores sensíveis, como IPs, domínios, nomes de instituições e pessoas, devem ser substituídos por pseudônimos consistentes via funções de hash criptográficas. A aplicação dessa pseudoanonimização precisa ser determinística (produzindo sempre o mesmo pseudônimo para o mesmo identificador) e reutilizável entre diferentes instâncias da ferramenta, possibilitando análises federadas e colaborativas entre CSIRTs distintos, sem violar a privacidade das fontes. O compartilhamento de informações sobre fontes de violações é uma meta perseguida por CSIRTs. O simples conhecimento da existência de um evento (ataque) em larga escala é valioso para CSIRTs de coordenação.

**(R2) Preservação do contexto operacional.** A anonimização não deve comprometer a estrutura dos relatos, preservando a cronologia, os vetores de ataque e as ações mitigadoras. A solução precisa manter o contexto semântico e sintático, substituindo apenas os dados necessários, sem afetar a compreensão do incidente, sua sequência ou gravidade.

**(R3) Suporte a dados não estruturados em múltiplos idiomas.** Considerando que *tickets* de incidentes são muitas vezes enviados por *e-mail*, formulários livres, ou até mesmo anexos, o sistema deve suportar minimamente textos não estruturados em Português e Inglês. Isso inclui o reconhecimento de entidades nominais técnicas (nomes de sistemas, IPs, domínios) e linguagem informal, com sensibilidade a erros ortográficos e siglas. Como o uso de imagens ou anexos de arquivos não textuais em *tickets* de segurança não é usual, entende-se que não é necessário incluí-lo como um requisito.

**(R4) Controle local de reidentificação e auditoria.** Embora a anonimização deva ser efetiva, alguns contextos, como auditorias ou investigações conjuntas, exigem reversibilidade. A solução deve manter localmente um método de reversibilidade (ex.*rainbow table*) com controle de acesso e trilha de auditoria, permitindo a reidentificação apenas por usuários autorizados. Também é desejável permitir o compartilhamento de dados com instituições parceiras, como CSIRTs de coordenação, para fins estatísticos.

**(R5) Interoperabilidade com ferramentas e padrões do ecossistema.** Os dados anonimizados devem manter compatibilidade com formatos de intercâmbio já utilizados por

CSIRTs, como o IODEF e sistemas de *tickets* estruturados (ex.: SGIS, RTIR, TOPdesk). Isto permite a integração com sistemas de orquestração, SIEMs, bancos de dados e futuras análises automatizadas. A solução a ser desenvolvida deve ter a capacidade de detectar essas estruturas para complementar as informações não estruturadas.

**(R6) Conformidade com LGPD.** É necessário manter a conformidade com a LGPD durante o processo de anonimização.

**(R7) Facilidade de uso e customização por equipes técnicas.** A solução deve permitir fácil configuração de novas entidades a anonimizar (ex.: via *regex* ou listas) e integração com fluxos de trabalho já utilizados por analistas. Isso inclui suporte a execuções em lote, geração de relatórios automatizados e interface amigável com sistemas de arquivos locais. A identificação de falhas de anonimização deve ser passível de sinalização para que um humano possa avaliar (*human-in-the-loop*).

**(R8) Conformidade com Regras locais.** O processo de anonimização deve ser realizado localmente, evitando o envio de dados para serviços externos de nuvem. Isso garante conformidade com normas internas institucionais, como em [UFRGS 2022]. Eventualmente, essas normas podem exigir adequação na anonimização e compartilhamento de dados sensíveis com terceiros.

A Tabela 3 apresenta um resumo dos requisitos para anonimização em CSIRTs, categorizando-os em requisitos funcionais (RF) e não funcionais (RNF). Os requisitos funcionais (R1-R4) focam em aspectos como a pseudoanonimização determinística, preservação do contexto, suporte a múltiplos idiomas e controle de reidentificação. Já os requisitos não funcionais (R5-R7) abordam a interoperabilidade com outras ferramentas, a execução local para conformidade com a LGPD e a facilidade de uso e customização.

| ID | Tipo | Descrição |
|---|---|---|
| R1 | RF | Pseudoanonimização determinística e reutilizável com consistência global |
| R2 | RF | Preservação do contexto técnico e semântico dos relatos |
| R3 | RF | Suporte a textos livres em múltiplos idiomas (PT/EN) |
| R4 | RF | Reidentificação controlada com registro de auditoria local |
| R5 | RNF | Compatibilidade com IODEF e sistemas de tickets de CSIRTs |
| R6 | RNF | Execução local e *compliance* com LGPD |
| R7 | RNF | Facilidade de customização e integração com fluxos existentes |

**Tabela 3. Requisitos Funcionais (RFs) e Não Funcionais (RNFs)**

## 5. *Framework* de anonimização de incidentes de segurança

A presente seção detalha a arquitetura do *framework* proposto e apresenta as tecnologias utilizadas na implementação.

### 5.1. Arquitetura

A arquitetura do *framework*, denominado de **`AnonLFI`**, foi projetada como um *pipeline* sequencial (conforme ilustrado na Figura 1) para garantir a anonimização eficaz e rastreável de dados sensíveis presentes em relatos de incidentes de segurança, tais como

os que podem ser vistos na Figura 2. A solução combina técnicas tradicionais, como expressões regulares, com abordagens modernas de NLP, permitindo a identificação precisa de padrões técnicos e reconhecimento de entidades nomeadas (NER).

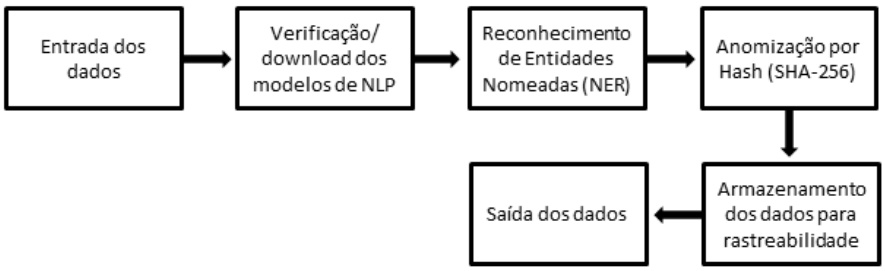



**Figura 1. Etapas do *pipeline* de processamento dos incidentes de segurança**

Cada uma das etapas do *pipeline* sequencial é detalhada a seguir:

1. **Entrada de Dados:** A ferramenta lê os dados de entrada a serem anonimizados, definido pelo usuário como único argumento posicional e obrigatório via linha de comando. É feita a identificação do formato (textual ou tabular) para que o processamento ocorra de forma adequada. Há suporte para arquivos `TXT`, `DOCX`, `CSV` e `XLSX`.
2. **Gerenciador de Modelos NLP:** O sistema verifica a existência local dos modelos específicos: '`pt_core_news_lg`', do spaCy; e o *transformer* '`Davlan/xlm-roberta-base-ner-hrl`'. O download do primeiro é realizado usando a API nativa do spaCy, enquanto o segundo é obtido do repositório Hugging Face[2]. Caso os arquivos dos modelos não estejam presentes no diretório local predefinido ('`models`'), o sistema realiza automaticamente o download e armazena para uso posterior. Essa etapa foi incorporada como medida de usabilidade, permitindo que o *pipeline* funcione de forma autônoma e sem necessidade de configurações manuais adicionais.
3. **Reconhecimento de Entidades Nomeadas (NER):** Utilizando o modelo *transformer* baixado e as regras RegEx, a ferramenta identifica entidades como nomes, organizações, localizações e endereços IP.
4. **Anonimização por Hash Criptográfica:** As detecções são substituídas por '*slugs*': o tipo da entidade, seguido por um fragmento de 10 caracteres do hash SHA-256 do texto original (sem uso de *salt* ou *nonce*). É gerado o hash completo da entidade limpa (sem espaços em branco), porém o *output* contém apenas os primeiros 10 caracteres como sufixo, no formato '`[TIPO_HASH10CHARS]`'.
5. **Armazenamento dos dados:** Todas as entidades são armazenadas em um SQLite local, escolhido intencionalmente por sua leveza e integração direta, sem dependências externas. Grava-se o tipo da entidade, seu texto original, o pseudônimo gerado (*slug*), o SHA-256 completo (identificador único) e *timestamps* de rastreabilidade (primeira e última ocorrência registrada).
6. **Saída e Relatórios:** Os dados anonimizados são salvos em um diretório predefinido ('`output/`'), enquanto os relatórios de execução são armazenados separadamente em '`logs/`'.

[2]`https://huggingface.co/Davlan/xlm-roberta-base-ner-hrl`

O *pipeline* tem início com a entrada dos incidentes. Em seguida, realiza-se, quando necessário, o *download* e o carregamento automático dos modelos de NLP. As entidades identificadas são, então, substituídas por pseudônimos gerados por meio de funções de hash criptográficas[3], preservando-se a estrutura e o contexto semântico do texto. Os mapeamentos das anonimizações são armazenados em um banco de dados local, permitindo a rastreabilidade controlada por meio de consultas a esse banco. Como saída, o sistema gera os dados anonimizados, acompanhados de um relatório contendo métricas básicas da execução.

### 5.2. Tecnologias Utilizadas

A implementação do *framework* foi realizada em *Python* e, visando maior facilidade de reprodução, utilizou-se o gerenciador uv[4], que lida com todas as dependências necessárias, orquestrando a execução, a instalação das bibliotecas e o versionamento de todas as tecnologias empregadas[5].

Para o desenvolvimento das etapas 2, 3 e 4, foram escolhidos o *framework* Microsoft Presidio [Microsoft 2025], integrado ao spaCy [AI 2025] e ao Hugging Face [Face 2025]. O Presidio fornece as principais funcionalidades associadas à anonimização, enquanto o spaCy é responsável pelo processamento de linguagem natural. O reconhecimento das entidades sensíveis foi realizado com o auxílio do modelo `Davlan/xlm-roberta-base-ner-hrl`[6]. Por fim, utilizou-se também o SQLite3[7] para o armazenamento dos dados (etapa 5), assegurando a consistência na atribuição de pseudônimos e o controle das identificações (com *timestamps*). Todas as tecnologias empregadas possuem licenças permissivas (MIT ou Apache 2.0), com uso permitido em ambientes acadêmicos ou corporativos.

## 6. Avaliação Experimental

Os dados utilizados na avaliação experimental foram coletados a partir de dois CSIRTs distintos. Os centros forneceram conjuntos de arquivos em formatos variados, incluindo *tickets* estruturados provenientes de diferentes sistemas (como o formato IODEF), bem como dados não estruturados, como *e-mails* com anexos encaminhados por usuários finais. Após o processamento e consolidação dessas fontes heterogêneas, foi compilado um conjunto final composto por 763 relatos de incidentes de segurança.

A metodologia de avaliação empregada neste trabalho consistiu em uma análise experimental que utiliza abordagens quantitativas e qualitativas para verificar a eficácia do *framework* de anonimização proposto. Essa avaliação foi estruturada em três etapas distintas. Primeiramente, realizamos uma avaliação da capacidade de anonimização do *framework*, detalhada na Seção 6.1, com foco na efetividade em remover ou mascarar informações sensíveis. Em segundo lugar, conforme apresentado na Seção 6.2, foi conduzida uma análise quantitativa comparando o desempenho do *framework* na análise de incidentes individuais. A terceira análise realizada em Seção 6.3 compara a solução proposta com resultados de um estudo anterior que empregou o modelo GPT-4 com *feedback*

[3] `https://csrc.nist.gov/projects/hash-functions`

[4] `https://docs.astral.sh/uv/`

[5] Código e exemplos de saída: `https://anonymous.4open.science/r/anon-F0EB`.

[6] `https://huggingface.co/Davlan/xlm-roberta-base-ner-hrl`

[7] `https://www.sqlite.org`

*adversarial* em anonimização de dados pessoais. Nesta análise, foram utilizadas quatro métricas principais: preservação do contexto semântico, legibilidade do texto anonimizado, utilidade das informações anonimizáveis remanescentes e grau de personalização da ferramenta. Finalmente, na Seção 6.4, os requisitos iniciais de projeto do *framework* foram revisitados utilizando critérios de avaliação como a cobertura da anonimização e o grau de atendimento aos requisitos previamente estabelecidos na Seção 4.

Os experimentos práticos foram conduzidos em uma estação de trabalho com a seguinte configuração de *hardware*: 16GB de memória RAM DDR4, CPU AMD Ryzen 3 3300X (8 núcleos lógicos), GPU NVIDIA GeForce GTX 1650 e um disco SSD KINGSTON SNV2S1000G. Importante ressaltar que não foram utilizados recursos de aceleração por GPU (CUDA ou similar) nem paralelização através de módulos de multiprocessamento durante a execução dos experimentos.

### 6.1. Efetividade na Anonimização de Incidentes de Segurança

Durante a fase de desenvolvimento, diferentes conjuntos de dados foram testados pela equipe, com a comparação entre os dados originais e os resultados da anonimização sendo avaliada de forma iterativa. As Figuras 2 e 3 ilustram um exemplo de incidente real e sua respectiva versão anonimizada. Nesse exemplo, observa-se que informações sensíveis como domínios e endereços IP foram corretamente anonimizadas. No entanto, casos em que esse processo falhou – como na linha "`From: Fail2Ban (Keyweb AG)`", onde o trecho "`(Keyweb AG)`" não foi anonimizado – foram considerados erros. Por outro lado, elementos como nomes de usuário em logs de tentativas de acesso `ssh` (caso de "`Invalid user yi`"), não foram anonimizados por decisão de projeto do sistema.

Com o objetivo de obter uma avaliação externa independente para esta publicação, três analistas de um CSIRT participaram da revisão manual dos 763 *tickets* anonimizados, com base em sua experiência na área de tratamento de incidentes, a fim de identificar possíveis falhas no processo de anonimização. Um dos analistas realizou a revisão detalhada dos textos anonimizados, enquanto os demais colaboraram com a classificação dos incidentes. Considera-se que o refinamento da anonimização é uma tarefa contínua, essencial para manter a qualidade dos resultados diante de situações não contempladas pelos datasets utilizados. Uma das limitações observadas refere-se a relatos de incidentes que incluem imagens com informações sensíveis embutidas no conteúdo visual, as quais não são processadas pela versão atual da ferramenta e não estão cobertas pelos *datasets* utilizados.

No conjunto de dados entregue aos analista, foram detectadas e anonimizadas 6.100 instâncias de entidades sensíveis (um endereço IP recorrente é contado uma única instância). Um dos analistas identificou 164 ocorrências de falsos negativos—ou seja, casos em que informações sensíveis deixaram de ser anonimizadas—, resultando em uma taxa de sucesso de 97,38% do processo de anonimização. O analista não identificou nenhum falso-positivo, *i.e.,* algum conteúdo não sensível que tenha sido anonimizado.

Um aspecto relevante observado nos casos de falso-negativo foi a concentração das 164 ocorrências em apenas 18 entidades distintas, indicando que a maioria dos erros estava associada a um subconjunto reduzido de padrões, muitos deles associados a nomes de domínio (DNS). Outro ponto importante é que 104 relatos de incidente estavam redigidos em idiomas estrangeiros (sendo 103 em Inglês e 1 em Alemão). Além disso,

diversos incidentes continham trechos em mais de um idioma, totalizando 412 segmentos multilíngues. Esse cenário corroborou com o requisito de necessidade de suporte multilíngue no processo de anonimização—um *script* complementar foi implementado para identificar automaticamente os trechos em Inglês.

Embora a análise externa tenha demonstrado a existência de falhas pontuais, estas podem ser corrigidas por meio da criação de regras adicionais ou ajustes nos modelos, sem comprometer, em nenhum momento, os requisitos e a arquitetura proposta.

Trechos de um Incidente Real (IPv4 alterados segundo RFC 3330)

```
From: Fail2Ban (Keyweb AG) <fail2ban-no-reply@dns01.keymachine.de>
To: mail-abuse@cert.br,cert@cert.br
Subject: Abuse from 192.0.2.138
Date: Tue, 22 Mar 2022 18:21:22 +0100 (CET)
Message-ID: <20220322172122.6B47FAE037D@dns01.keymachine.de>

Dear Sir/Madam,
We have detected abuse from the IP address ( 192.0.2.138 ), which according to a whois lookup is on
your network. We would appreciate if you would investigate and take action as appropriate.
Any feedback is welcome but not mandatory.
Log lines are given below, but please ask if you require any further information.

==================== Excerpt from log for 192.0.2.138 ====================
Note: Local timezone is +0100 (CET)
Mar 22 17:38:57 dns01 sshd[3082249]: Invalid user yi from 192.0.2.138 port 51570
Mar 22 17:38:57 dns01 sshd[3082249]: pam_unix(sshd:auth): authentication failure; logname=
uid=0 euid=0 tty=ssh ruser= rhost=192.0.2.138
Mar 22 17:38:59 dns01 sshd[3082249]: Failed password for invalid user yi from 192.0.2.138 port
51570 ssh2
Mar 22 17:44:59 dns01 sshd[3082984]: Invalid user testing from 192.0.2.138 port 59998
```

**Figura 2. Exemplo de relato de incidente recebido por email**

Incidente pós-anonimização

```
From: [ORGANIZATION_dc2075266f] (Keyweb AG) <[EMAIL_ADDRESS_5de3f9a416]>
To: [EMAIL_ADDRESS_c4122f8a60],[EMAIL_ADDRESS_83824c64b2]
Subject: Abuse from [IP_ADDRESS_542860d50d]
Date: Tue, 22 Mar 2022 18:21:22 +0100 (CET)
Message-ID: <[EMAIL_ADDRESS_32161ebd15]>

Dear Sir/Madam,
We have detected abuse from the IP address ( [IP_ADDRESS_542860d50d] ), which according to a whois
lookup is on your network. We would appreciate if you would investigate and take action as appropriate.
Any feedback is welcome but not mandatory.
Log lines are given below, but please ask if you require any further information.

==================== Excerpt from log for [IP_ADDRESS_542860d50d] ====================
Note: Local timezone is +0100 (CET)
Mar 22 17:38:57 dns01 sshd[3082249]: Invalid user yi from [IP_ADDRESS_542860d50d]
port 51570
Mar 22 17:38:57 dns01 sshd[3082249]: pam_unix(sshd:auth): authentication failure; logname=
uid=0 euid=0 tty=ssh ruser= rhost=[IP_ADDRESS_542860d50d]
Mar 22 17:38:59 dns01 sshd[3082249]: Failed password for invalid user yi from [IP_ADDRESS_542860d50d]
port 51570 ssh2
Mar 22 17:44:59 dns01 sshd[3082984]: Invalid user testing from [IP_ADDRESS_542860d50d] port 59998
```

**Figura 3. Exemplo de relato de incidente anonimizado com o `AnonLFI`**

Apesar de não terem sido identificadas ocorrências de falsos-positivos na amostra analisada, reconhece-se que esse tipo de erro pode ocorrer em contextos mais amplos, especialmente em domínios ambíguos, como nomes comuns ou termos técnicos. Como estratégia de mitigação, a ferramenta pode ser configurada para adotar uma abordagem conservadora em cenários de incerteza, optando pela anonimização preventiva nesses casos. Tal abordagem busca priorizar a proteção de dados sensíveis, ainda que à custa de uma leve perda de fidelidade no conteúdo, especialmente em contextos nos quais a anonimização serve como etapa de pré-processamento para LLMs.

De modo geral, a eficácia da ferramenta—**com uma taxa de 97,38% de sucesso**—é bastante satisfatória, especialmente considerando a diversidade linguística presente no conjunto de teste e o fato de o sistema ter como requisito operar exclusivamente de forma local (SLM). Os principais tipos de informação— endereços IP e e-mails—foram corretamente anonimizados quase na totalidade dos casos. Ainda assim, o sistema apresenta limitações ao lidar com padrões sintáticos mais complexos como um endereço IP sintaticamente incorreto, ou entidades com caracteres especiais (“|”, “(”, “)” e “_”).

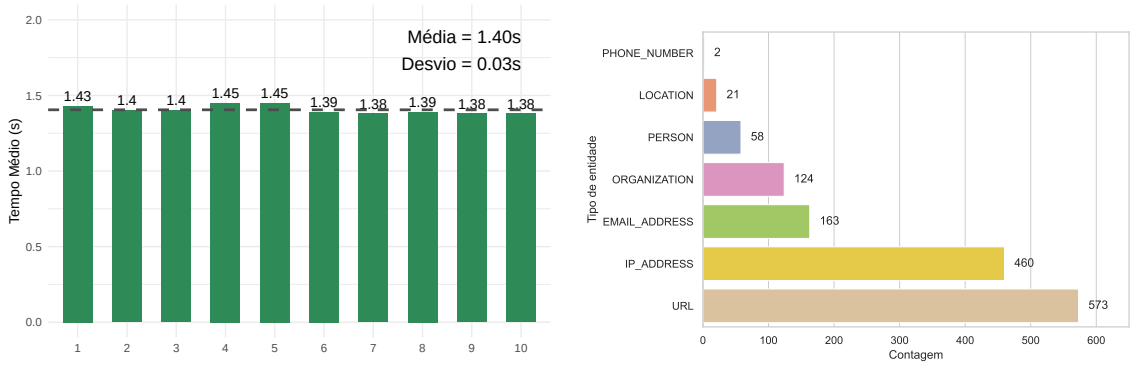
### 6.2. Avaliação Quantitativa

A avaliação quantitativa envolve a análise objetiva do desempenho da ferramenta, considerando tanto a eficiência quanto a abrangência na anonimização. Foram adotadas métricas como o número total de incidentes processados, o tempo médio de execução e a frequência de diferentes tipos de PII (*Personally Identifiable Information*[8]). A análise foi conduzida por meio de dez execuções consecutivas da ferramenta, orquestradas pelo *script* `get_runs_metrics.py`, disponível no repositório da ferramenta. Esse *script* percorre os arquivos de teste, aplica a anonimização e extrai estatísticas consolidadas, incluindo o tempo médio por execução e a quantidade de entidades detectadas por tipo.

Os resultados apresentados na Figura 4 evidenciam a estabilidade e a abrangência do *framework*. O primeiro gráfico (Figura 4a) mostra que o tempo médio de processamento por *ticket* permaneceu praticamente constante ao longo das dez execuções, com **média de 1,40 segundos por ticket** e desvio padrão de apenas 0,03 segundos, indicando consistência e previsibilidade no desempenho. O segundo gráfico (Figura 4b) mostra a distribuição das entidades sensíveis detectadas, com destaque para `URL`, `IP_ADDRESS` e `EMAIL_ADDRESS`, o que está em conformidade com os elementos mais recorrentes nos relatos de incidentes analisados. Esses resultados reforçam a capacidade da ferramenta em lidar com os principais tipos de PII presentes em ambientes operacionais.

Um tempo médio de 1,4 segundos por incidente é considerado satisfatório, especialmente considerando que os dados são processados localmente e que um CSIRT de porte médio costuma receber menos de 100 *tickets* por dia. Ainda assim, cabe destacar que há espaço para otimizações de desempenho, uma vez que o suporte à execução com GPU ainda não foi implementado, por exemplo.

[8] `https://csrc.nist.gov/glossary/term/personally_identifiable_information`




**(a) Tempo médio por ticket** **(b) Contagem de tipos de entidades**

**Figura 4. Média de Tempo por Ticket e Contagem de Entidades Únicas**

### 6.3. Avaliação Qualitativa

Como forma de complementar as métricas quantitativas de desempenho da anonimização, realizamos também uma comparação com os resultados apresentados por [Staab et al. 2024], que utilizaram o modelo GPT-4 com *feedback adversarial* (GPT-4AA) para anonimização de dados pessoais. Utilizamos gráficos para visualizar o desempenho em quatro métricas qualitativas: (1) preservação do contexto semântico, (2) legibilidade, (3) utilidade das informações anonimizáveis e (4) grau de personalização da ferramenta.

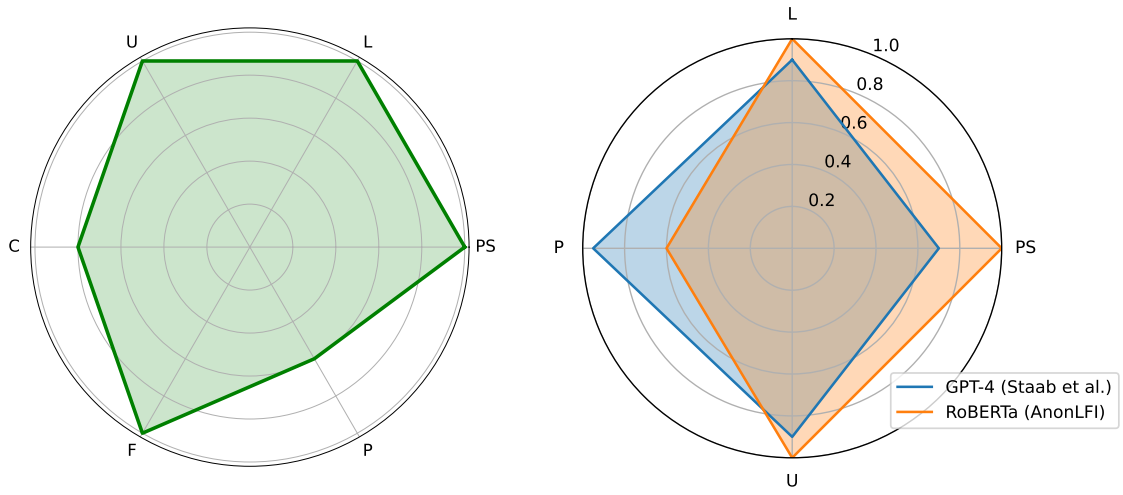
Os valores de referência (GPT-4AA) foram extraídos da Tabela 1 do artigo de [Staab et al. 2024], expressos em uma escala de 0 a 1. Já os nossos resultados, obtidos com o modelo RoBERTa, foram originalmente avaliados em uma escala de 0 a 5 e, posteriormente, normalizados para a mesma escala de comparação. A seguir, apresentamos a análise individual para cada critério.

- **Preservação do contexto semântico (PS):** a anonimização substitui SAs (*sensitive attributes*) por *placeholders* descritivos (como *EMAIL*, *IP_ADDRESS*, *ORGANIZATION*), de forma a manter o sentido original das frases.
- **Legibilidade (L):** o texto anonimizado mantém fluidez e coesão gramatical, sem interrupções abruptas ou marcações excessivas (leitura natural).
- **Grau de personalização (P):** a ferramenta permite a definição de novas regras de entidades sensíveis a partir da modificação do código pela interface da API do Microsoft Presidio, além de integrar modelos de NLP de maneira modular, sendo esta feita pela mudança dos modelos no código, conforme demonstrado na Figura 5.

```
ALLOW_LIST = ["TCP", "UDP", "HTTP", "HTTPS", "admin", "localhost"]
TRANSFORMER_MODEL = "Davlan/xlm-roberta-base-ner-hrl"
```

**Figura 5. Personalização pela alteração de modelos de NLP em código**

- **Utilidade das informações anonimizáveis (U):** o conteúdo permanece útil para fins de correlação, investigação e treinamento de modelos.




**(a) AnonLFI (escala 0–5)** **(b) AnonLFI vs. Staab et al. (0–1)**

**Figura 6. Resumo da avaliação qualitativa**

Além das métricas utilizadas para comparação com o trabalho de [Staab et al. 2024], avaliamos também a **(C)obertura de anonimização** (nota 4) e **(F)acilidade de integração** (nota 5) como critérios adicionais de qualidade da solução. Estas métricas foram consideradas por sua relevância prática: a primeira se refere à capacidade da ferramenta de anonimizar de forma consistente diferentes tipos de informações sensíveis (não restritas a), enquanto a segunda diz respeito à facilidade de incorporar a solução em fluxos de trabalho ou sistemas já existentes. Ambas foram avaliadas de maneira subjetiva e independente, sem correspondência direta nas métricas dos trabalhos relacionados. As Figuras 6a e 6b ilustram que nossa solução apresenta desempenho comparável ou superior em quase todas as métricas avaliadas.

### 6.4. Critérios de Validação e Revisão dos Requisitos Iniciais

Por fim, foi realizada uma avaliação adicional com foco no atendimento aos requisitos definidos, por meio de um conjunto estruturado de critérios. No contexto da Engenharia de Software, critérios são condições objetivas e verificáveis utilizadas para avaliar ou aceitar artefatos, atividades ou decisões, assegurando a conformidade com requisitos, padrões ou objetivos previamente estabelecidos [ISO/IEC/IEEE 2018]. Com base nos oito requisitos originais (Tabela 3), foram definidos doze critérios de validação. A Tabela 4 apresenta a correspondência entre eles e a avaliação do `AnonLFI` em cada critério.

Os resultados da avaliação demonstram que todos os oito critérios estabelecidos foram atendidos de forma satisfatória, com destaque para os critérios C1 e C3, cujos resultados foram muito próximos ou iguais a 100%. Os critérios C1, C3, C10 e C12 foram validados a partir da avaliação externa conduzida por um analista de CSIRT. O critério C4 foi avaliado com base no *dataset* disponível, embora ajustes possam ser necessários futuramente, considerando a possível inclusão de novos objetos contendo dados sensíveis, como arquivos de áudio ou vídeo. O critério C8 é considerado parcialmente atendido, pois é necessário que o SIEM utilizado ofereça suporte à exportação de dados ou API de integração direta e contínua, em tempo real. Além disso, testes mais amplos de integração, com uma variedade maior de SIEMs e outras ferramentas, poderão revelar ajustes ou incrementos futuros no *framework* proposto.

| Critério | Descrição resumida | AnonLFI | Requisito |
|---|---|---|---|
| C1 | Precisão na identificação e anonimização de atributos sensíveis | ✓ (97%) | R1, R2 |
| C2 | Suporte a múltiplas técnicas de anonimização (RegEx, hash, NER) | ✓ | R1, R2 |
| C3 | Redação seletiva que preserva o conteúdo não sensível | ✓ (100%) | R2 |
| C4 | Suporte a dados não estruturados, como logs, e-mails e relatórios | ✓ | R3 |
| C5 | Reconhecimento e anonimização multilíngue com precisão (PT/EN) | ✓ | R3 |
| C6 | Reversibilidade controlada com acesso restrito (ex. rainbow table) | ✓ | R4 |
| C7 | Geração de logs e rastreabilidade para auditoria e conformidade | ✓ | R4, R6 |
| C8 | Interoperabilidade com SIEMs, bancos de dados e ferramentas de monitoramento (se prover API ou exportação de dados) | △ Parcial | R5 |
| C9 | Licenciamento aberto que permite uso, modificação e redistribuição | ✓ MIT | R6 |
| C10 | Uso *offline* e conformidade com LGPD e políticas institucionais | ✓ | R6, R8 |
| C11 | Linguagem de desenvolvimento compatível e de fácil customização | ✓ Python | R7 |
| C12 | Facilidade de instalação, configuração e manutenção da ferramenta | ✓ | R7 |

**Tabela 4. Correspondência entre critérios de validação, avaliação do AnonLFI e requisitos propostos**

## 7. Conclusão

Ao longo deste trabalho, foram analisadas diversas técnicas e ferramentas de anonimização de dados, com ênfase em incidentes de segurança. Discutiram-se abordagens como pseudoanonimização, uso de funções *hash*, e a aplicação de *frameworks* modulares, destacando o equilíbrio entre a proteção da privacidade e a manutenção da utilidade dos dados. A pesquisa levou em conta a validação rigorosa das transformações realizadas e a atenção aos desafios operacionais, como vulnerabilidades e limitações técnicas, compondo uma análise crítica e abrangente das soluções disponíveis.

Como resultado prático, foi desenvolvido o **AnonLFI**, um *framework* híbrido e modular para anonimização de dados sensíveis, como endereços IP, e-mails e *hostnames*. Utilizando expressões regulares, técnicas de NLP e persistência em banco de dados, o sistema viabiliza rastreabilidade controlada, conciliando anonimização eficaz com a capacidade de correlação de eventos. Entre suas limitações, destacam-se a falha na detecção de fragmentos de IP (ex.: três octetos), problemas na identificação de endereços IPv6, atribuídos a um *bug* na biblioteca Presidio [Github 2024] e a diversidade de *datasets*.

Como trabalhos futuros, propõe-se a integração do *framework* a sistemas SIEM, como *Splunk* e *ElasticSearch*, bem como a investigação do uso de criptografia homomórfica para suportar o compartilhamento seguro de bases anonimizadas de incidentes. Além disso, sugere-se uma avaliação sistemática da resiliência do *framework* frente a ataques de reidentificação conduzidos por modelos de linguagem adversariais—capazes de inferir entidades sensíveis a partir de padrões linguísticos e contextuais preservados após o processo de anonimização.

## References

AI, E. (2025). spacy: Industrial-strength natural language processing. https://spacy.io/.

Aleroud, A., Yang, F., Pallaprolu, S. C., Chen, Z., and Karabatis, G. (2021). Anonymization of network traces data through condensation-based differential privacy. *Digital Threats*, 2(4).

ANT Lab, I. (2018). cryptopant ip address anonymization library. https://ant.isi.edu/software/cryptopANT/index.html.

Face, H. (2025). Transformers documentation.

Fejrskov, M., Pedersen, J. M., and Vasilomanolakis, E. (2020). Cyber-security research by isps: A netflow and dns anonymization policy. In *Cyber Security*, pages 1–8.

Fisk, G., Ardi, C., Pickett, N., Heidemann, J., Fisk, M., and Papadopoulos, C. (2015). Privacy principles for sharing cyber security data. In *IEEE S&P*, pages 193–197. IEEE.

Github (2024). Ip recognizer has bugs when ipv6 contains double colon '::' · issue #1476 · microsoft/presidio.

Gunay, M., Keles, B., and Hizlan, R. (2024). Llms-in-the-loop part 2: Expert small ai models for anonymization and de-identification of phi across multiple languages.

Haber, A. C., Sax, U., and Prasser, F. (2022). Open tools for quantitative anonymization of tabular phenotype data: literature review. *Briefings in Bioinformatics*, 23.

Imperva (2025). What is data anonymization | pros, cons & common techniques. https://www.imperva.com/learn/data-security/anonymization/.

ISO/IEC/IEEE (2018). ISO/IEC/IEEE 29148:2018 - Systems and software engineering — Life cycle processes — Requirements engineering.

Koukis, D., Antonatos, S., Antoniades, D., Markatos, E., and Trimintzios, P. (2006). A generic anonymization framework for network traffic. In *IEEE ECC*, pages 2302–2309.

Majeed, A. and Lee, S. (2020). Anonymization techniques for privacy preserving data publishing: A comprehensive survey. *IEEE access*, 9:8512–8545.

Majeed, A. and Lee, S. (2021). Anonymization techniques for privacy preserving data publishing: A comprehensive survey. *IEEE Access*, 9:8512–8545.

Microsoft (2025). Presidio: Open-source pii anonymization and detection. https://microsoft.github.io/presidio/.

Murthy, S., Bakar, A. A., Rahim, F. A., and Ramli, R. (2019). A comparative study of data anonymization techniques. In *IEEE BigDataSecurity*, pages 306–309. IEEE.

Plonka, D. (2003). ip2anonip. https://pages.cs.wisc.edu/ plonka/ip2anonip.

Portillo-Dominguez, A. O. and Ayala-Rivera, V. Towards an efficient log data protection in software systems through data minimization and anonymization. In *2019 7th CONISOFT*.

Prasser, F., Kohlmayer, F., Spengler, H., and Kuhn, K. A. (2017). ARX - a comprehensive tool for anonymizing biomedical data. *AMIA Annual Symposium Proceedings*.

Rasic, A. (2020). Anonymization of event logs for network security monitoring. Master's thesis, Concordia University. Unpublished.

Senavirathne, N. and Torra, V. (2020). On the role of data anonymization in machine learning privacy. In *IEEE TrustCom*, pages 664–675.

Slagell, A., Lakkaraju, K., and Luo, K. (2006). Flaim: A multi-level anonymization framework for computer and network logs.

Staab, R., Vero, M., Balunović, M., and Vechev, M. (2024). Large language models are advanced anonymizers. *arXiv preprint arXiv:2402.13846*.

Tempest (2021). Desmistificando a anonimização de dados | sidechannel. https://www.sidechannel.blog/anonimizacao-de-dados-o-que-para-que-e-como-e/.

Templ, M., Kowarik, A., and Meindl, B. (2015). Statistical disclosure control for micro-data using the R package sdcMicro. *Journal of Statistical Software*, 67(4):1–36.

UFRGS (2022). A inteligência artificial generativa e a proteção dos dados pessoais, da privacidade e da propriedade intelectual.

Varanda, A., Santos, L., de C. Costa, R. L., Oliveira, A., and Rabadão, C. (2021). Log pseudonymization: Privacy maintenance in practice. *Journal of Information Security and Applications*, 63:103021.

Wiest, I. C., Wolf, F., Leßmann, M.-E., van Treeck, M., Ferber, D., Zhu, J., Boehme, H., Bressem, K. K., Ulrich, H., Ebert, M. P., et al. (2024). Llm-aix: An open source pipeline for information extraction from unstructured medical text based on privacy preserving large language models. *medRxiv*.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., and Brew, J. (2019). Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

Xu, J., Fan, J., Ammar, M., and Moon, S. B. (2001). On the design and performance of prefix-preserving ip traffic trace anonymization. In *ACM SIGCOMM IMW*.

Yan, B., Li, K., Xu, M., Dong, Y., Zhang, Y., Ren, Z., and Cheng, X. (2024). On protecting the data privacy of large language models (llms): A survey.

Yang, L., Tian, M., Xin, D., Cheng, Q., and Zheng, J. (2024). AI-driven anonymization: Protecting personal data privacy while leveraging machine learning.

Zhang, J., Borisov, N., and Yurcik, W. (2006). Outsourcing security analysis with anonymized logs. In *2006 Securecomm and Workshops*, pages 1–9.