Mascaramento por Agrupamento e Rotulagem com LLMs para Compartilhamento de Datasets de Incidentes em Redes

Breno Valente Manhães[1], **Guilherme A. Thomaz**[1], **Miguel Elias M. Campista**[1]

[1]Grupo de Teleinformática e Automação
Universidade Federal do Rio de Janeiro (UFRJ)

{bvalente,guiaraujo,miguel}@gta.ufrj.br

Abstract. *The dissemination of network security datasets is often limited by sensitive attributes in textual logs generated by tools such as OpenVAS and Nessus. This paper proposes the MECAL (Masking via Embedding Clustering and Automated Labeling) algorithm to anonymize these attributes while preserving their utility. The method utilizes Transformers to semantically cluster incident descriptions and employs Large Language Models (LLMs) to generate high-level generic labels for each cluster. Results demonstrate that replacing the original texts with the generated labels improves data quality, evidenced by improvements in F1-Score and Mutual Information metrics, enabling the secure sharing of cyber defense information.*

Resumo. *A disseminação de bases de dados de segurança de redes é frequentemente limitada por atributos sensíveis presentes em logs textuais de ferramentas como OpenVAS e Nessus. Este trabalho propõe o algoritmo MECAL (Mascaramento por Clusterização de Embeddings e Rotulagem Automática) para anonimizar esses atributos preservando sua utilidade. O método utiliza Transformers para agrupar semanticamente as descrições de incidentes e emprega LLMs (Large Language Models) para gerar rótulos genéricos de alto nível para cada grupo. Os resultados demonstram que a substituição dos textos originais por rótulos gerados melhora a qualidade dos dados, como evidenciado pelo aumento das métricas de F1-Score e Mutual Information, viabilizando o compartilhamento seguro de informações de defesa cibernética.*

1. Introdução

A aplicação de técnicas de Aprendizado de Máquina (*Machine Learning* – ML) e Inteligência Artificial (IA) tem se tornado fundamental para a defesa cibernética, permitindo a detecção de intrusões e a análise de vulnerabilidades com maior rapidez e precisão. No entanto, o desenvolvimento e a validação desses modelos dependem intrinsecamente da disponibilidade de bases de dados (*datasets*) representativas e atualizadas [Ring et al. 2019]. Apesar da existência de repositórios públicos, a comunidade científica e a indústria enfrentam uma escassez crítica de dados reais de incidentes, motivada principalmente por questões de confidencialidade e de regulação.

Este trabalho foi realizado com apoio da Rede Nacional de Ensino e Pesquisa (RNP), da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) – Código de Financiamento 001 e 88887.954253/2024-00, do CNPq (310234/2025-5, 407304/2025-8, 408255/2023-4 e 405940/2022-0), da FAPERJ (E-26/200.380/2023 e E-26/210.778/2025), da FAPESP (2023/00673-7 e 2023/00811-0) e da Fundação de Desenvolvimento da Pesquisa - Fundep - Rota 2030 em conjunto dos nossos parceiros Stellantis e Mobway. Agradecimentos especiais aos Profs. Igor Monteiro Moraes e Reinaldo Cézar de M. Gomes.

Organizações relutam em compartilhar seus logs de segurança devido ao risco de expor infraestruturas críticas ou violar legislações de proteção de dados, como a LGPD (Lei Geral de Proteção de Dados) no Brasil e a GDPR (Regulamento Geral de Proteção de Dados) na Europa. Estudos indicam que barreiras legais e a falta de garantias de privacidade são os principais impedimentos à troca efetiva de inteligência sobre ameaças cibernéticas [Wagner et al. 2019]. Ferramentas de varredura de vulnerabilidades, como OpenVAS e Nessus, geram relatórios ricos em descrições textuais não estruturadas. Embora essenciais para a compreensão do contexto de um incidente, esses campos de texto livre frequentemente contêm informações sensíveis, como endereços IP privados, versões específicas de software, nomes de usuários e caminhos de diretórios, que atuam como identificadores diretos ou quase-identificadores da organização-alvo.

As técnicas tradicionais de anonimização, como a supressão de colunas ou o mascaramento sintático (e.g., substituição por *hashes*), tendem a remover a utilidade semântica dos dados, inviabilizando análises qualitativas subsequentes. Por outro lado, a generalização manual de textos, que dependeria da curadoria humana para abstrair informações detalhadas em rótulos de alto nível, é um processo oneroso e pouco escalável diante do volume massivo de logs gerados diariamente [Deußer et al. 2025]. Existe, portanto, uma lacuna tecnológica para métodos que consigam automatizar a transformação de dados textuais sensíveis em representações seguras, porém semanticamente úteis.

Este trabalho propõe o algoritmo MECAL (Mascaramento por Clusterização de *Embeddings* e Rotulagem Automática), uma abordagem híbrida que combina técnicas de Processamento de Linguagem Natural (PLN) baseadas em redes *transformers* com o poder de abstração de Modelos de Linguagem de Grande Escala (*Large Language Models* – LLMs). O método converte descrições textuais de vulnerabilidades em vetores semânticos (*embeddings*), agrupa-os por similaridade de contexto e utiliza um LLM para gerar rótulos sintéticos de alto nível para cada grupo. O resultado é uma base de dados na qual informações sensíveis originais são substituídas por categorias genéricas padronizadas, preservando a distribuição estatística e o significado operacional dos incidentes. Portanto, as principais contribuições deste trabalho são: (i) o desenvolvimento do algoritmo MECAL, focado na anonimização de atributos textuais não estruturados em logs de segurança; (ii) uma metodologia de rotulagem automatizada via LLM que elimina a necessidade de curadoria manual de grupos; e (iii) a avaliação da qualidade dos dados transformados, demonstrando que a abordagem preserva melhor a estrutura de correlação dos dados do que técnicas tradicionais de supressão.

Os experimentos realizados validam as três contribuições, evidenciando uma redução de mais de 65% na entropia de Shannon e uma queda na razão de unicidade para 0,09%, garantindo robustez contra reidentificação. Em termos de utilidade, o MECAL superou a supressão de dados na detecção de anomalias (*F1-score* de 0,75) e manteve alta fidelidade à estrutura semântica original (informação mútua normalizada de 0,84).

Este trabalho está organizado da seguinte forma: a Seção 2 discute trabalhos relacionados sobre anonimização de dados e bases de dados de segurança. A Seção 3 aborda as limitações das técnicas tradicionais. A Seção 4 detalha a arquitetura e o funcionamento do algoritmo MECAL. A Seção 5 discorre sobre as métricas e os experimentos realizados para avaliar a usabilidade e a preservação dos dados. Por fim, as avaliações experimentais e as conclusões são apresentadas nas Seções 6 e 7, respectivamente.

2. Trabalhos Relacionados

A anonimização de logs de segurança situa-se na interseção entre o processamento de linguagem natural (PLN), a privacidade de dados e a cibersegurança. Como consequência, a análise da literatura recente divide-se em três eixos principais: técnicas de reconhecimento de entidades nomeadas (*Named Entities Recognition* – NER), microagregação semântica e o uso de LLMs para sanitização.

2.1. Limitações de Abordagens Baseadas em Entidades (NER)

Historicamente, a sanitização de logs dependeu de expressões regulares e listas de bloqueio para remover identificadores diretos (*Personally Identifiable Information* – PII), como endereços IP e nomes de usuários. Com o avanço do aprendizado profundo, *frameworks* como o SDLog [Zhang and Li 2025] passaram a utilizar modelos contextualizados (e.g., CodeBERT) para identificar informações sensíveis com alta precisão. Contudo, essas abordagens operam majoritariamente no nível de *"span"* (trechos de texto), substituindo termos específicos por *placeholders*.

No contexto de relatórios de vulnerabilidades (OpenVAS/Nessus), a simples remoção de entidades nomeadas é insuficiente. Machado et al. propuseram recentemente o uso de LLMs para a extração estruturada de dados desses relatórios, visando a normalizar dados provenientes de formatos heterogêneos [Machado et al. 2025]. Embora avancem na estruturação, os autores reconhecem que a anonimização permanece um desafio, pois a combinação de descrições técnicas, mesmo sem endereços IP, pode formar uma "impressão digital" da infraestrutura (não-anonimizabilidade epistêmica), permitindo a reidentificação da organização-alvo através de vetores de ataque específicos.

2.2. Microagregação Semântica

Para mitigar os riscos de inferência, a literatura tem migrado para a microagregação, onde registros são agrupados para satisfazer a propriedade de k-anonimato. Aufschläger et al. introduziram o *ClustEm4Ano*, um método que utiliza *embeddings* para gerar hierarquias de generalização automáticas para atributos textuais [Aufschläger et al. 2024]. Embora o *ClustEm4Ano* represente um avanço significativo ao eliminar a necessidade de taxonomias manuais, ele foi projetado primordialmente para atributos nominais curtos (e.g., "Ocupação", "País"). A aplicação direta em textos longos e de alta entropia, como logs de erros, resulta na perda de nuances técnicas críticas. Além disso, métodos tradicionais de microagregação textual, como o M-MDAV [Garg and Torra 2023], focam na substituição de textos por centroides numéricos ou geométricos, o que, apesar de preservar a estatística, destrói a interpretabilidade humana do registro anonimizado.

2.3. LLMs para Sanitização

A geração de dados sintéticos ou sanitizados via LLMs é a tendência mais recente. O sistema INTACT [Pilán et al. 2025] propõe uma sanitização guiada por "ataques de inferência", na qual um LLM reescreve documentos substituindo termos sensíveis por generalizações verazes (hiperônimos). Paralelamente, a abordagem de *Summaries as Centroids* [Grootendorst 2025] sugere o uso de LLMs para criar representações textuais de grupos.

O algoritmo MECAL, proposto neste trabalho, difere dessas abordagens ao integrar a robustez estatística da microagregação à capacidade interpretativa dos LLMs. Diferentemente do INTACT, que altera palavras no texto (arriscando vazamentos contextuais), o MECAL substitui o registro inteiro por um rótulo de classe gerado que oferece uma síntese aproximada do conteúdo do grupo. Essa estratégia alinha-se a soluções industriais, como o sistema Clio da Anthropic [Anthropic 2024], que agrega dados de conversação em "facetas" de alto nível para proteger a privacidade individual, mas adapta esse conceito especificamente para o detalhamento técnico exigido em operações de defesa cibernética.

3. Técnicas Tradicionais de Transformação e Anonimização de Atributos

A anonimização de dados estruturados é fundamentada em conceitos como o k-anonimato (*k-anonymity*) [Sweeney 2002], que visa garantir que cada registro em uma base de dados tenha atributos indistinguíveis de pelo menos $k - 1$ outros registros. Para alcançar tal propriedade em bases de dados relacionais, técnicas tradicionais incluem a supressão (remoção completa de atributos sensíveis) e a generalização (substituição de valores específicos por intervalos ou categorias mais amplas, e.g., substituir uma idade "24" pelo intervalo "20-30"). Porém, o tratamento de atributos de linguagem natural não estruturados, como os campos de descrição e de solução em relatórios de vulnerabilidades, apresenta desafios únicos que essas técnicas clássicas não conseguem abordar satisfatoriamente.

3.1. Desafio da alta cardinalidade e variabilidade textual

Diferentemente de atributos categóricos fixos (e.g., "Protocolo: TCP/UDP"), as descrições de incidentes são altamente variáveis. Uma simples mudança na versão de um software gera uma string única. A aplicação de técnicas de pseudonimização, como o *hashing* criptográfico (e.g., SHA-256), embora eficaz para mascarar o valor original, destrói a interpretabilidade dos dados. Como funções de *hash* são projetadas para evitar colisões, textos semanticamente idênticos com variações mínimas de sintaxe resultam em *hashes* completamente diferentes, impedindo que algoritmos de aprendizado de máquina identifiquem padrões ou similaridades entre incidentes relacionados.

3.2. Limitações do One-Hot Encoding

Para viabilizar o processamento de textos em modelos de aprendizado de máquina, é necessário converter os dados em representações numéricas. A técnica de transformação mais comum para dados categóricos é o *One-Hot Encoding* (OHE). O OHE cria uma nova coluna binária para cada valor único do atributo original. Embora robusto para categorias com baixa cardinalidade, o OHE torna-se inviável para textos de logs de segurança devido a dois problemas fundamentais: explosão de dimensionalidade e esparsidade, e perda de relacionamento semântico.

A explosão de dimensionalidade e esparsidade é uma consequência de textos de vulnerabilidades longos e variados, sendo que o número de valores únicos tende a ser muito próximo ou idêntico ao número de registros. A aplicação do OHE resultaria em uma matriz com milhares de colunas, onde a vasta maioria dos valores seria zero. Isso introduz a "maldição da dimensionalidade" [Cerda et al. 2018], aumentando drasticamente o custo computacional e exigindo volumes de dados muito maiores para evitar o sobreajuste dos modelos. Já a perda de relacionamento semântico é uma consequência do OHE, já que cada valor único é tratado como ortogonal aos demais. Isso significa que, sob a

ótica do algoritmo, uma vulnerabilidade descrita como "*Buffer Overflow in Apache 2.4*" seria considerada tão distinta de "*Buffer Overflow in Apache 2.5*" quanto seria de "*SQL Injection*". A representação vetorial binária não captura a proximidade semântica entre os termos, ignorando que as duas primeiras falhas pertencem à mesma classe de problemas.

Essas limitações levam a adoção de abordagens que, não apenas mascarem o dado bruto, mas que também reduzam a dimensionalidade preservando a semântica do incidente, motivando o uso de *embeddings* e agrupamentos, como proposto neste trabalho.

4. Algoritmo Proposto: MECAL

Esta seção apresenta o algoritmo Mascaramento por Clusterização de *Embeddings* e Rotulagem Automática (MECAL). A proposta visa solucionar o desafio de anonimizar campos de texto livre e não estruturados, especificamente descrições de vulnerabilidades e soluções recomendadas, sem a perda total de utilidade analítica que ocorreria com a simples supressão dos dados. O MECAL opera como uma técnica de generalização semântica automatizada, transformando textos de alta entropia em rótulos categóricos discretos através de um fluxo que combina aprendizado de máquina não supervisionado e a capacidade de abstração dos LLMs.

4.1. Arquitetura e fluxo de execução

O processo de transformação dos dados é ilustrado na Figura 1, considerando apenas seis amostras textuais para simplificar a visualização. O fluxo MECAL é dividido em três etapas sequenciais, detalhadas a seguir.

A Etapa 1 consiste na conversão do texto bruto (T) em um vetor numérico denso, ou *embedding* (V), representados pelos retângulos vermelhos na Figura 1. Para isso, utiliza-se a biblioteca *SentenceTransformers* baseada na arquitetura BERT [Reimers and Gurevych 2019]. Especificamente, este trabalho adota o modelo pré-treinado `all-MiniLM-L6-v2` [Hugging Face 2021]. A escolha deste modelo justifica-se pelo equilíbrio entre desempenho computacional e qualidade semântica. O modelo mapeia sentenças para um espaço vetorial de 384 dimensões, no qual textos com significados similares são posicionados em proximidade geométrica (alta similaridade de cosseno), independentemente da sobreposição exata de palavras-chave. O processamento é acelerado via GPU, usando o *framework* NVIDIA CUDA.

Na Etapa 2, aplica-se o algoritmo *K-means* para discretizar o espaço vetorial contínuo gerado na etapa anterior. O objetivo é identificar similaridades nas descrições de vulnerabilidades e agrupá-las em conjuntos semanticamente coesos. Definiu-se empiricamente o hiperparâmetro $K = 20$ grupos, devido à eficiência na segregação de tipos distintos de vulnerabilidades[1]. Para garantir a reprodutibilidade dos experimentos, o estado aleatório foi fixado (`random_state=0`). A localização dos vetores de *embedding* no espaço de dimensão 384 é representada na Figura 1 através de apenas três dimensões ou eixos, e os grupos (*clusters*) são representados por elipses.

A Etapa 3 consiste na rotulagem generativa automatizada. A ideia é atribuir um rótulo semântico único e descritivo para cada um dos grupos identificados pelo *K-means*. Com isso, o MECAL se diferencia de abordagens que exigem curadoria manual intensiva

[1]Este trabalho utiliza K maiúsculo para o número de grupos e k minúsculo para o nível de k-anonimato.

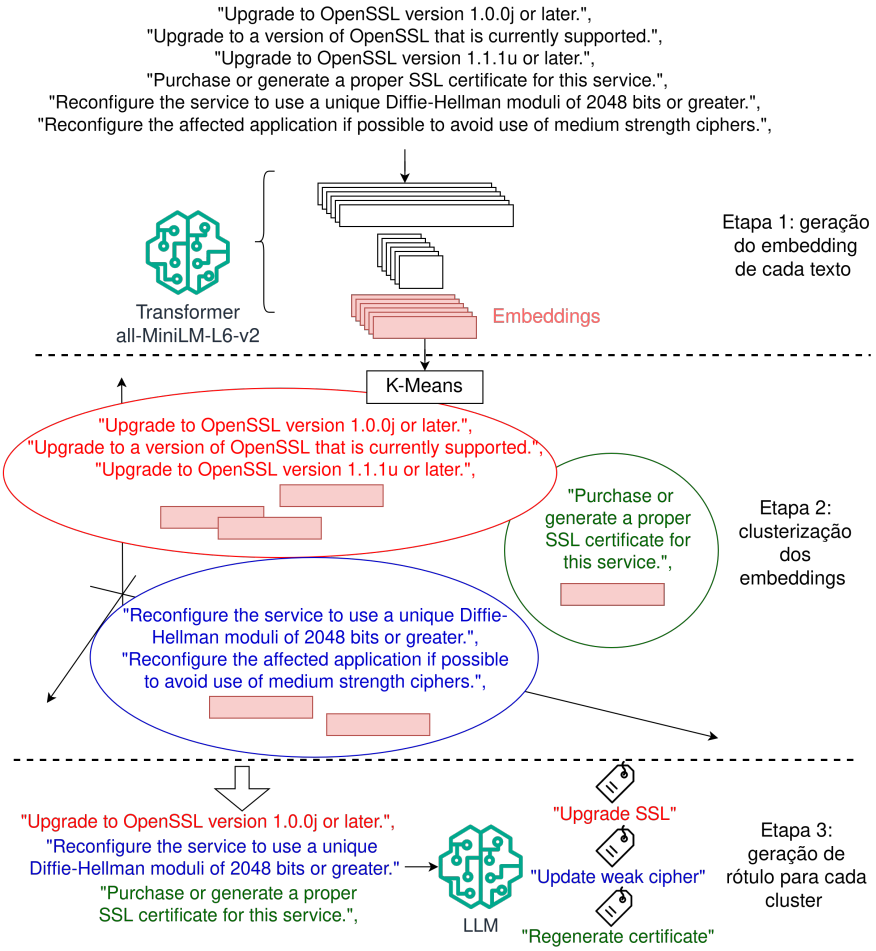

Figura 1. Fluxo de processamento do algoritmo MECAL , de cima para baixo.

ou fusão de grupos. O processo ocorre da seguinte forma. Para cada grupo C_i: (i) cinco amostras representativas do grupo (os textos cujos *embeddings* estão mais próximos do centroide geométrico de C_i) são selecionadas; (ii) estas cinco amostras são enviadas como contexto em um *prompt* para um LLM; e (iii) o LLM é instruído a analisar os padrões comuns nas amostras e gerar um rótulo curto e representativo que sintetize o tópico daquele grupo específico. Na Figura 1, apenas uma amostra por grupo é selecionada para simplificar a ilustração, e os rótulos textuais de cada grupo são identificados pelas cores correspondentes. Essa abordagem garante escalabilidade e consistência, eliminando a necessidade de fusão manual de grupos. Ressalta-se que, embora cada grupo seja processado individualmente, o MECAL permite que múltiplos grupos distintos ($C_i \neq C_j$) sejam mapeados para o mesmo rótulo textual, caso o LLM identifique que ambos representam o mesmo conceito semântico subjacente. Assim, o número final de classes únicas pode ser $\leq K$, promovendo uma consolidação natural dos dados. A Tabela 1 exemplifica este comportamento, onde grupos diferentes convergem para o mesmo rótulo.

Na etapa final, os textos originais das colunas `descrição` e `solução`, que frequentemente contêm versões de software, caminhos de diretórios e IPs privados, são descartados. Em seu lugar, são inseridas as colunas `classe` e uma nova coluna `tipo_solução`, contendo apenas os rótulos gerados automaticamente na etapa anterior.

Tabela 1. Exemplo de mapeamento: Grupos de descrição para rótulos gerados por LLM. Os grupos 1 e 5 possuem descrições muito similares, gerando o mesmo rótulo.

ID do Grupo	Rótulo Gerado pelo LLM (classe)
0	Outdated SSL/TLS Libraries
1	**Buffer Overflow in Kernel**
2	Weak Ciphers Configuration
3	Expired Certificates
4	Driver Compatibility Issues
5	**Buffer Overflow in Kernel**
...	...

4.2. Complexidade Computacional e Reprodutibilidade

A complexidade assintótica do MECAL é $\mathcal{O}(nK)$, resultante da composição de suas três etapas: a geração de *embeddings*, linear em n; o agrupamento por K-Means, que escala com $O(nKdi)$ para dimensão d e iterações i fixas, reduzindo-se a $O(nK)$; e a rotulagem via LLM, que realiza K chamadas de custo constante. O termo dominante é o do K-Means, pois $nK > n$ para qualquer $K > 1$. Para ilustrar a viabilidade em escala real, estima-se um tempo de execução inferior a seis minutos para $n = 100.000$ registros e $K = 100$ grupos, valores muito acima dos utilizados neste estudo, sem exigir infraestrutura computacional especializada além de uma GPU de uso geral.

A implementação do algoritmo e os *scripts* de avaliação foram disponibilizados publicamente em um repositório online[2], permitindo a verificação experimental desses resultados. Devido às restrições de confidencialidade da RNP, o conjunto de dados original não pôde ser disponibilizado no repositório.

4.3. Mecanismo de Mitigação de Ataques de Referência Cruzada

A arquitetura do MECAL foi projetada não apenas para garantir o k-anonimato, mas também para oferecer resistência estrutural a ataques de inferência, em especial os de referência cruzada. Esse vetor de ameaça ocorre quando um adversário utiliza conhecimento externo ou correlações determinísticas presentes nos dados para deduzir valores de atributos sensíveis que foram suprimidos [Fung et al. 2010]. Em logs de segurança em que descrições textuais ricas são mantidas, mesmo quando identificadores explícitos (como CVEs ou IPs) são removidos, a existência de uma bijeção (relação um-para-um) entre o texto e o atributo oculto compromete a anonimização.

A justificativa para a abordagem baseada em agrupamento semântico fundamenta-se na existência de padrões reais de correlação identificados na base de dados utilizada neste trabalho (vide Seção 5.1). Observa-se, por exemplo, o caso de uma vulnerabilidade crítica real presente nos registros (doravante referenciada como **CVE-AAAA-NNNN**, CVSS 9.8). Em abordagens tradicionais de anonimização, o texto de solução associado, contendo instruções específicas como "`Upgrade to [Software X] version [v.v.v]...`", mapeia exclusivamente para esta CVE. Consequentemente, a presença desse texto atua como um quase-identificador forte: um atacante com acesso à base transformada poderia cruzar o texto da solução com bases públicas (e.g.,

[2]`https://github.com/GTA-UFRJ/MECAL`

National Vulnerability Database – NVD) e inferir, com total confiança, a presença da vulnerabilidade crítica na infraestrutura alvo.

O MECAL neutraliza esse vetor de ataque ao romper a correlação determinística através da generalização. Ao substituir os textos originais pelos rótulos de classe gerados pelo LLM, o algoritmo aproxima-se do conceito de l-diversidade [Machanavajjhala et al. 2007]. Instruções específicas de versionamento são abstraídas para rótulos genéricos, como "`Atualização de Software`" ou "`Gerenciamento de Patches`". Desta forma, um único rótulo gerado pelo MECAL passa a englobar não apenas a CVE-AAAA-NNNN, mas também um conjunto diversificado de outras vulnerabilidades ($l \geq 2$) com características semânticas similares, mas que afetam softwares ou versões distintas. Essa característica do algoritmo introduz incerteza no nível do atributo textual: ao observar o rótulo de saída, torna-se inviável determinar qual vulnerabilidade específica gerou o registro original. Assim, o processo de mascaramento proposto bloqueia a inferência direta enquanto preserva a utilidade semântica necessária para análises estatísticas de categorias de incidentes.

5. Metodologia

Para validar a eficácia do MECAL, adotou-se uma avaliação dual, analisando o compromisso entre a proteção de dados sensíveis (**privacidade**) e a manutenção da coerência semântica dos dados (**utilidade**). Os experimentos foram conduzidos sobre uma base de dados real de incidentes, focando na transformação dos atributos textuais `descrição` e `solução` para os seus respectivos rótulos gerados, `classe` e `tipo_solução`.

5.1. Base de Dados

Os experimentos foram realizados utilizando uma base de dados real de incidentes de segurança de redes, cedida pela Rede Nacional de Ensino e Pesquisa (RNP). O conjunto de dados é composto por 15.187 registros de vulnerabilidades gerados automaticamente pela ferramenta OpenVAS (*Open Vulnerability Assessment System*). A escolha dessa base de dados é estratégica devido à natureza não estruturada e técnica dos registros gerados por ferramentas de varredura (*scanners*). Cada registro contém atributos textuais de alta entropia, especificamente as colunas `descrição` (descrição técnica da vulnerabilidade) e `solução` (solução recomendada), além de metadados estruturados como severidade (*Common Vulnerability Scoring System* – CVSS) e endereços de rede. O foco dos experimentos recai sobre a transformação e anonimização desses atributos textuais para os seus respectivos rótulos gerados, `classe` e `tipo_solução`.

5.2. Análise qualitativa

Para demonstrar a superioridade do agrupamento semântico proposto pelo MECAL em relação às técnicas convencionais de extração de palavras-chave, realizou-se uma análise comparativa qualitativa. O objetivo é evidenciar que métodos baseados puramente em frequência de termos ou extração de frases não capturam adequadamente o *significado* das descrições de vulnerabilidades, resultando em agrupamentos incoerentes. Para isso, foram avaliadas três técnicas alternativas aplicadas a uma amostra de 500 descrições. A primeira técnica consiste na vetorização das descrições utilizando o *Term Frequency-Inverse Document Frequency* (TF-IDF), seguida de agrupamento via *K-means* com $K = 20$ grupos,

em que os rótulos são gerados automaticamente a partir das palavras com maior peso TF-IDF em cada grupo. A segunda abordagem utiliza o algoritmo RAKE (*Rapid Automatic Keyword Extraction*) [Rose et al. 2010] para extrair a frase-chave mais relevante de cada descrição, agrupando-as posteriormente por palavras-chave idênticas. Por fim, utilizou-se um LLM (Llama 3.1 8B) para gerar uma única palavra-chave por descrição, seguida do agrupamento por termos idênticos. A avaliação qualitativa consiste em examinar amostras de cada grupo formado por esses métodos e verificar se as descrições compartilham coerência semântica, ou seja, se tratam efetivamente do mesmo tipo de vulnerabilidade ou problema de segurança.

5.3. Métricas de privacidade

A avaliação de privacidade visa quantificar a redução do risco de reidentificação e a eliminação de especificidades técnicas que poderiam atuar como quase-identificadores. Foram empregadas as seguintes métricas:

Redução de Entropia de Shannon: A Entropia de Shannon (H) mede o grau de incerteza ou a quantidade média de informação contida em uma variável aleatória [Shannon 1948], $H(X) = -\sum_{i=1}^{n} P(x_i) \log_2 P(x_i)$, onde $P(x_i)$ é a probabilidade de ocorrência do valor x_i. O resultado é a quantidade de bits necessária para codificar cada amostra como um número binário diferente. Por ser uma média entre as amostras, o valor em bits pode ser fracionário. Quanto mais uniforme for a distribuição dos dados, menor é a entropia. Em logs de segurança, descrições em texto livre possuem alta entropia devido à variabilidade de caracteres e detalhes técnicos. O objetivo do MECAL é reduzir essa entropia, generalizando descrições únicas para categorias padronizadas. A redução é calculada comparando-se a entropia do texto original (X) e a do rótulo gerado (Y). Uma redução significativa na entropia indica que o algoritmo condensou com sucesso a dispersão dos dados brutos em um conjunto finito de conceitos.

Informação Mútua (*Mutual Information* – MI): A Informação Mútua quantifica a dependência entre duas variáveis, i.e., quanta informação sobre o texto original (X) é preservada no rótulo anonimizado (Y), i.e., $I(X;Y) = H(X) - H(X|Y)$. Neste contexto, a MI é utilizada para calcular a privacidade preservada. Assume-se que a informação que *não* é capturada pela MI entre o original e o rótulo representa a "perda de informação" benéfica (remoção de ruído, endereços IP e versões específicas), que contribui para a privacidade.

Redução de Cardinalidade: a cardinalidade é o número de valores distintos. A eficiência do mascaramento é medida pela transformação de um espaço de milhares de descrições distintas em um conjunto compacto de categorias.

Unicidade: mede a proporção de registros que possuem valores distintos na base de dados. Registros únicos ($k = 1$) são vulneráveis a ataques de ligação (*linkage attacks*).

Análise de Classes de Equivalência (k-anonimato): Para verificar a conformidade com o princípio de k-anonimato [Sweeney 2002], analisa-se a distribuição de tamanho dos grupos gerados. Cada rótulo produzido pelo MECAL define uma classe de equivalência. A métrica crítica aqui é o tamanho do menor grupo ($|C|_{min}$), que determina o valor de k da base de dados transformado, garantindo que cada incidente seja indistinguível de, pelo menos, $k - 1$ outros incidentes dentro do mesmo grupo semântico.

5.4. Métricas de Utilidade

A avaliação da utilidade dos dados anonimizados foi conduzida através de dois experimentos principais, desenhados para medir a preservação das propriedades estatísticas e semânticas do conjunto de dados original após a aplicação do MECAL.

O primeiro experimento avalia a capacidade do conjunto de dados transformado de manter a distribuição original. Caso isso ocorra, o processo de anonimização preserva a capacidade de diferenciar eventos de segurança comuns (maior probabilidade) e raros (menor probabilidade). Para testar essa hipótese, foi desenvolvido um *pipeline* de processamento que termina com a pontuação de cada amostra de acordo com um escore de chance de ocorrência, que inclusive permite identificar pontos muito incomuns (*outliers*). Primeiramente, aplica-se uma PCA (*Principal Component Analysis*) para manter apenas as 25 componentes principais que aumentam a variância explicada dos dados, mitigando a maldição da dimensionalidade. Em seguida, uma KDE (*Kernel Density Estimation*) estima a função de densidade de probabilidade multidimensional dos dados, gerando um valor de chance de ocorrência (*likelihood*) p_i para cada amostra. Por fim, os valores de $L_i = \log p_i$ são normalizados, subtraindo-se a média e dividindo-se pelo desvio padrão, resultando no escore final de chance de ocorrência. O resultado com o MECHE é comparado a um *baseline* em que as colunas de descrição e solução são codificadas com *One-Hot*. A avaliação deu-se com o *F1-Score*, que mede a sobreposição entre os top 5% *outliers* detectados no *baseline* e nas versões anonimizadas, e pelo RMSE (*Root Mean Square Error*), que quantifica a divergência entre as curvas de escores de anomalia.

Já o segundo experimento analisa o impacto na clusterização, para verificar a preservação da estrutura semântica das vulnerabilidades, utilizando o algoritmo HDBSCAN (*Hierarchical Density-Based Spatial Clustering of Applications with Noise*). O objetivo foi avaliar a aplicação de agrupamento de vulnerabilidades para atribuição a equipes de segurança. Foram comparados três cenários distintos. O Cenário A (NLP) utilizou dados originais enriquecidos com *embeddings* semânticos gerados pelo modelo `all-MiniLM-L6-v2`, representando o limite superior de qualidade, no qual os *embeddings* não são agrupados. O Cenário B utilizou as colunas textuais geradas pelo algoritmo MECAL. Por fim, o Cenário C (supressão) considerou apenas vetores CVSS e escores numéricos, sem informação textual ou categórica derivada. A qualidade intrínseca dos grupos resultantes foi medida via *silhouette score* e Validação de Clusterização Baseada em Densidade (*Density-Based Clusterization Validation* – DBCV). A fidelidade semântica foi avaliada pelo cálculo da Informação Mútua Normalizada (*Normalized Mutual Information* – NMI) entre os *clusters* gerados pelo MECAL (caso B) e os do NLP (Caso A, *baseline*), sendo que um alto valor de NMI indica que a informação semântica fundamental é preservada mesmo com a alteração das fronteiras exatas dos grupos.

6. Resultados

Esta seção demonstra quantitativamente como a substituição dos textos livres (`descrição` e `solução`) pelos rótulos gerados (`classe` e `tipo_solução`) impacta as propriedades de privacidade e a estrutura estatística dos dados, como visto na Seção 5.

6.1. Análise qualitativa

A análise comparativa revelou diferenças substanciais na qualidade dos agrupamentos produzidos por cada técnica. A Tabela 2 resume os principais problemas identificados.

O método TF-IDF gerou rótulos como "04 / lts / ubuntu", agrupando descrições por palavras frequentes de texto padronizado e não pelo tipo de vulnerabilidade – -o mesmo grupo continha falhas em sistemas de arquivos, servidores web e kernel. O RAKE extraiu frases do texto padrão dos avisos (e.g., "*preceding description block directly*"), resultando em 50% das amostras na categoria residual "outros". O LLM gerou palavras-chave semanticamente relevantes, porém sem consolidação: variações como "*kernel vulnerability*" e "*kernel vulnerabilities*" permaneceram em grupos distintos, fragmentando 40% dos dados. O MECAL superou essas limitações ao combinar *embeddings* semânticos com agrupamento e rotulagem. Grupos como "Biblioteca criptográfica" (110 itens) e "Pacote de kernel/driver" (67 itens) demonstram coerência semântica, pois o modelo `all-MiniLM-L6-v2` captura o significado contextual, permitindo agrupar vulnerabilidades descritas com vocabulário diferente, mas semanticamente relacionadas.

Tabela 2. Comparação qualitativa das técnicas de agrupamento.

Técnica	Grupos	Problema Principal
TF-IDF	20	Rótulos são frequências de palavras, não conceitos semânticos
RAKE	31	50% dos dados em "outros"; captura *boilerplate* de advisories
LLM	70	Alta fragmentação (70 palavras-chave para 100 amostras)
MECAL	**14**	**Grupos semânticos coerentes com rótulos interpretáveis**

6.2. Análise de privacidade e redução de identificabilidade

A principal contribuição do MECAL para a privacidade é a transformação de atributos quase-identificadores únicos em classes de equivalência densas. A Tabela 3 resume a comparação das métricas de entropia, unicidade e cardinalidade antes e depois da aplicação do algoritmo. A tabela mostra uma queda abrupta na entropia para o campo de descrição, partindo de 9,96 bits para 3,46 bits (redução de 65,23%), indicando que o MECAL removeu com sucesso a variabilidade excessiva que tornava cada registro "único". A métrica mais impactante para a proteção contra ataques de reidentificação é a redução da cardinalidade. Originalmente, a base de dados possuía 2.722 descrições únicas. O MECAL mapeou todo esse universo para apenas 14 classes semânticas (queda de 99,48%). A razão de unicidade (*uniqueness ratio*) caiu de aproximadamente 18% para 0,09% no campo de descrição, indicando uma queda de 17,83 pontos percentuais (p.p.). Isso elimina efetivamente o risco de ataques de referência cruzada como os descritos na Seção 4.3, nos quais um atacante isola um alvo baseando-se na "assinatura" textual específica de uma vulnerabilidade. Após o processamento, um registro não é mais identificável por sua descrição técnica, mas sim indistinguível de todos os outros membros da mesma classe. Os resultados para o atributo de solução são análogos, mas a diversidade inicial dos textos é inferior, resultando em uma menor redução nas métricas.

Já a análise da distribuição dos grupos confirma que o MECAL impõe um alto grau de k-anonimato (Tabela 4). Para o atributo `descrição`, o menor grupo formado ("SNMP") agrupa 143 registros. Isso implica que, considerando apenas este atributo, a base de dados satisfaz $k_{min} = 143$. Ou seja, qualquer incidente anonimizado é indistinguível de pelo menos outros 142 incidentes. A média de tamanho das classes foi de aproximadamente 194 registros, demonstrando uma distribuição equilibrada que evita a criação de grupos muito pequenos e vulneráveis. No atributo `solução`, embora a dispersão original fosse menor, o MECAL garantiu um tamanho mínimo de grupo de 35 re-

Tabela 3. Comparativo de Métricas de Privacidade: Original vs. MECAL.

Métrica	Atributo: Descrição → Classe			Atributo: Solução → Tipo_Solução		
	Original	**MECAL**	**Redução**	**Original**	**MECAL**	**Redução**
Entropia de Shannon	9,96	3,46	**65,23%**	5,92	3,16	**46,66%**
Valores Únicos (Cardinalidade)	2.722	14	**99,48%**	946	20	**97,88%**
Razão de Unicidade	17,92%	0,09%	**17,83 p.p.**	6,23%	0,13%	**6,1 p.p.**

gistros (para a categoria "*Configure HTTP policies and HSTS*"), com um tamanho médio de classe de 47,3 registros.

Tabela 4. Estatísticas das Classes de Equivalência Geradas (Grupos).

Estatística	**Descrição → Classe**	**Solução → Tipo_Solução**
Total de grupos	14	20
Tamanho médio dos grupos	194,4	47,3
Tamanho do menor grupo (k_{min})	143	35
Tamanho do maior grupo	690	240

A informação mútua (MI) entre os textos originais e os rótulos gerados revela o balanço entre generalização e fidelidade. Para as descrições, a MI relativa foi de 24,1%. Isso deve ser interpretado positivamente no contexto de privacidade: significa que aproximadamente 75,9% da informação contida no texto bruto, correspondente a detalhes específicos, identificadores e variações sintáticas, foi intencionalmente suprimida, retendo-se apenas o "núcleo" semântico necessário para a categorização (os 24,1% restantes). Resultados similares foram observados para as soluções, com 63% de "privacidade preservada" (informação removida) e 37% de retenção semântica. Esse resultado reflete o fato de as soluções originais serem menos aleatórias (apresentarem menor entropia), característica que reduz intrinsecamente a informação mútua calculada.

6.3. Análise de utilidade

Os experimentos de utilidade demonstram que a preservação da semântica através das categorias do MECAL oferece vantagens mensuráveis em comparação à simples supressão dos dados textuais. A Tabela 5 apresenta os resultados da detecção de anomalias baseada em escore de chance de ocorrência, descrita na Seção 5.4. O MECAL alcançou uma *F1-Score* de 0,75, superando a abordagem de supressão que obteve 0,70. Isso indica que o MECAL preserva a distribuição de probabilidades, o que permite identificar vulnerabilidades raras, comportando-se de maneira mais similar à base de dados original completa. A raiz do erro médio quadrático (RMSE) do MECAL também foi inferior (0,42 contra 0,46), confirmando que a distribuição dos escores de chance de ocorrência gerados se aproxima mais da distribuição de referência do que a versão suprimida.

Tabela 5. Desempenho na Detecção de Anomalias (Referência: Texto Completo).

Método	**F1-Score**	**RMSE**	**Anomalias em Comum**
MECAL	**0,7511**	**0,4262**	504
Supressão	0,7094	0,4614	476

Na avaliação de agrupamento, conforme detalhado na Tabela 6, observa-se um fenômeno interessante no qual a remoção de atributos textuais complexos aumenta as métricas de coesão interna dos grupos (*silhouette score* e DBCV). Tanto o MECAL quanto a supressão apresentaram valores de *silhouette score* superiores a 0,82, enquanto o uso dos textos completos obteve 0,75. Esse aumento ocorre porque a redução da dimensionalidade simplifica o espaço de dados, facilitando a formação de agrupamentos densos baseados puramente em atributos numéricos de risco. Entretanto, essas métricas, baseadas na estrutura espacial dos dados (qualidade dos grupos), não capturam a preservação da semântica. A análise de fidelidade semântica revela a distinção crucial entre o MECAL e a supressão. A informação mútua normalizada (NMI) de 0,84 entre os grupos formados pelo MECAL e os grupos formados pelo texto original indica uma altíssima correlação informativa. As fronteiras exatas dos grupos mudaram significativamente, refletido pelo baixo Índice de Rand Ajustado (*Adjusted Rand Index* – ARI) de 0,29. Entretanto, a NMI elevada comprova que o MECAL não destrói a estrutura semântica dos dados, mas sim a reorganiza de forma consistente. O algoritmo preserva a utilidade analítica ao manter as relações fundamentais entre os tipos de vulnerabilidade, algo que a supressão total de dados não é capaz de garantir.

Tabela 6. Qualidade de Agrupamento e Fidelidade Semântica.

Cenário	Qualidade do Grupo		Fidelidade Semântica	
	DBCV	***Silhouette score***	**ARI**	**NMI**
A: *Baseline* (texto completo)	0,43	0,75	1,00	1,00
B: MECAL	**0,52**	0,82	0,29	**0,84**
C: Supressão	0,52	**0,83**	-	-

7. Conclusões

Este trabalho apresentou o algoritmo MECAL, uma abordagem inovadora para o mascaramento de registros de incidentes de segurança baseada no agrupamento de *embeddings* e rotulagem automática via LLM. Os resultados demonstram que a substituição de descrições textuais muito diversas (alta entropia) por rótulos semânticos resumidos permite o compartilhamento de bases de dados de rede em conformidade com requisitos de privacidade, sem inviabilizar a análise estatística e operacional dos dados. Como trabalhos futuros, pretende-se conduzir um estudo sistemático do comportamento das métricas de privacidade e utilidade em função do número de grupos K, substituindo a escolha empírica atual por um critério baseado em evidência. Adicionalmente, planeja-se avaliar o MECAL em *datasets* de segurança publicamente disponíveis, a fim de verificar a capacidade de generalização da abordagem para diferentes padrões de logs de rede. Por fim, pretende-se ampliar a comparação do MECAL com outros métodos da literatura.

Referências

Anthropic (2024). Clio: Privacy-preserving insights into real-world ai use. Technical report, Anthropic Research.

Aufschläger, R., Wilhelm, S., Heigl, M., and Schramm, M. (2024). Clustem4ano: Clustering text embeddings of nominal textual attributes for microdata anonymization. In *Proceedings of the 28th International Database Engineering & Applications Symposium (IDEAS)*. arXiv:2412.12649.

Cerda, P., Varoquaux, G., and Kégl, B. (2018). Similarity encoding for learning with dirty categorical variables. *Machine Learning*, 107(8):1477–1494.

Deußer, T., Sparrenberg, L., Berger, A., Hahnbück, M., Bauckhage, C., and Sifa, R. (2025). A survey on current trends and recent advances in text anonymization. *arXiv preprint arXiv:2508.21587*.

Fung, B. C., Wang, K., Chen, R., and Yu, P. S. (2010). Privacy-preserving data publishing: A survey of recent developments. *ACM Computing Surveys (CSUR)*, 42(4):1–53.

Garg, S. and Torra, V. (2023). K-anonymous privacy preserving manifold learning. In *Proceedings of the 15th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K)*. Umeå University, SciTePress.

Grootendorst, M. (2025). Summaries as centroids for interpretable and scalable text clustering. *arXiv preprint arXiv:2502.09667*.

Hugging Face (2021). Hugging face model hub: all-minilm-l6-v2. `https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2`.

Machado, B., Lautert, D., Kapelinski, C., and Kreutz, D. (2025). Structured extraction of vulnerabilities in openvas and tenable was reports using llms. *arXiv preprint arXiv:2511.15745*.

Machanavajjhala, A., Kifer, D., Gehrke, J., and Venkitasubramaniam, M. (2007). l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):3:1–3:52.

Pilán, I., Manzanares-Salor, B., Sánchez, D., and Lison, P. (2025). Truthful text sanitization guided by inference attacks. *arXiv preprint arXiv:2412.12928*.

Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*.

Ring, M., Wunderlich, S., Scheuring, D., Landes, D., and Hotho, A. (2019). A survey of network-based intrusion detection data sets. *Computers & Security*, 86:147–167.

Rose, S., Engel, D., Cramer, N., and Cowley, W. (2010). Automatic keyword extraction from individual documents. In *Text Mining: Applications and Theory*, pages 1–20. John Wiley & Sons, Ltd.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570.

Wagner, T. D., Mahbub, K., Palomar, E., and Abdallah, A. E. (2019). Cyber threat intelligence sharing: Survey and research directions. *Computers & Security*, 87:101589.

Zhang, Y. and Li, X. (2025). Sdlog: A deep learning framework for detecting sensitive information in software logs. *arXiv preprint arXiv:2505.14976*.