CENTRO UNIVERSITÁRIO DO INSTITUTO MAUÁ DE TECNOLOGIA

Ciência da Computação

Data Exposure Score (DES) – Quantificando sua segurança

Carlos Henrique Lucena Barros

Débora Witkowski

Leonardo Cazotto Stuber

Mateus Capaldo Martins

Orientadores: Profa. Dra. Ana Paula Gonçalves Serra

Profa. Dra. Milkes Yone Alvarenga

Resumo

O presente trabalho propõe o *Data Exposure Score* (DES), um sistema destinado a mensurar o nível de exposição digital de indivíduos a partir de dados públicos compartilhados em redes sociais. O sistema foi implementado objetivando quantificar o risco de superexposição de informações pessoais e promover a conscientização sobre vulnerabilidades online, incentivando práticas mais seguras de uso das mídias digitais. A coleta de dados foi feita na rede social Bluesky, utilizando sua API aberta (AtProto) e scripts em Python para extração e armazenamento das postagens em Amazon DocumentDB. Uma análise comparativa automatizada foi conduzida por meio do Amazon Bedrock, que executou o modelo Llama 3.2 90B Instruct via API gerenciada e o modelo GPT-4o da OpenAI. O modelo interpreta o conteúdo textual e retorna uma estrutura JSON padronizada com valores booleanos que indicam a presença de informações sensíveis. Realizadas essas etapas, o DES calcula um escore ponderado conforme a criticidade e explorabilidade dos dados identificados, produzindo uma métrica numérica que reflete o grau de exposição digital de cada usuário. Os resultados obtidos apontam que o sistema demonstra a viabilidade de aplicar modelos LLM para a análise de segurança informacional, contribuindo para o debate sobre privacidade e educação digital.

Palavras-chave: Exposição Digital. Segurança da Informação. Redes Sociais. Inteligência Artificial. Análise de Risco

Abstract

This paper presents the development of the Data Exposure Score (DES), a system designed to measure individuals' digital exposure based on public data shared on social networks. Its goal is to quantify the risk of personal information overexposure and promote awareness of online vulnerabilities, encouraging safer digital practices. Data was collected from the Bluesky platform using its open API (AtProto) and processed through Python scripts, with storage in Amazon DocumentDB. A comparative analysis was performed via Amazon Bedrock, utilizing the Llama 3.2 90B Instruct model via a managed

API and the GPT-4o model from OpenAI. These models interpret textual content and return a standardized JSON structure with boolean values indicating the presence of sensitive information. Based on these results, the DES computes a weighted score according to the criticality and exploitability of the detected data, generating a numeric metric that reflects each user's digital exposure level. The system demonstrates the feasibility of applying large language models to information security analysis, contributing to the discussion on privacy and digital literacy.

Keywords: *Digital Exposure. Information Security. Social Media. Artificial Intelligence. Risk Analysis.*

1. Introdução

A era digital contemporânea é marcada pelo conceito de cibercultura, entendido como o modo de vida moldado pela presença constante das tecnologias digitais em praticamente todas as dimensões da sociedade — da comunicação ao entretenimento, dos negócios às interações pessoais (DICIO, 2017). Essa nova configuração social, impulsionada pela conectividade, trouxe avanços expressivos na forma como indivíduos se relacionam e compartilham informações, mas também estabeleceu um paradoxo essencial: quanto maior a interação e o engajamento *online*, maior tende a ser a extração de excedente comportamental e menor o controle sobre a privacidade dos dados pessoais (ZUBOFF, 2021).

As redes sociais representam, de maneira clara, essa dualidade. Criadas para potencializar a comunicação e fortalecer comunidades virtuais, são plataformas desenhadas para estimular o compartilhamento constante de informações, recompensando o comportamento de exposição por meio de curtidas, visualizações e validação social. Esse mecanismo, embora eficaz para engajamento, amplia significativamente a quantidade de dados sensíveis que os próprios usuários tornam públicos. Surge, assim, o fenômeno da autoexposição digital, no qual o indivíduo, muitas vezes de forma inconsciente, atua como agente do próprio vazamento de informações. Ao publicar dados sobre sua rotina, preferências, localização e identidade, ele contribui para formar uma vasta superfície de ataque explorável por terceiros mal-intencionados.

Esse problema ultrapassa o campo puramente técnico e se insere em uma dimensão sociotécnica complexa, na qual fatores humanos e tecnológicos interagem continuamente. O design das plataformas — baseado em algoritmos de recomendação, métricas de engajamento e estímulos comportamentais — combina-se ao comportamento espontâneo dos usuários, criando vulnerabilidades sistêmicas e um ambiente propício à exploração indevida de dados pessoais.

Diante desse cenário, o presente projeto propõe investigar a autoexposição de informações em redes sociais e apresentar o *Data Exposure Score* (DES), um sistema que busca mensurar e representar de forma objetiva o grau de exposição digital dos usuários a partir de suas próprias publicações. O sistema utiliza um escore, o DES, concebido como uma métrica inovadora, capaz de traduzir o risco digital em um valor numérico interpretável, permitindo ao indivíduo compreender, visualizar e reduzir sua vulnerabilidade informacional nas plataformas.

Este estudo se posiciona na convergência entre segurança da informação e análise de riscos cibernéticos, propondo uma abordagem prática e educativa para mitigar vulnerabilidades que emergem do uso cotidiano das redes. O DES, portanto, além de um sistema de medição, se estabelece como uma ferramenta de conscientização digital, destinada a promover hábitos mais seguros de interação online e a reduzir os impactos decorrentes do uso indevido de dados pessoais em ambientes virtuais.

1.1. Objetivos

Objetivo Geral

Desenvolver e implementar um sistema, o *Data Exposure Score* (DES), com um indicador (o escore DES) capaz de medir o grau de exposição digital de indivíduos a partir de dados extraídos de mídias, utilizando a rede social Bluesky como caso de estudo para quantificar a exposição.

Objetivos Específicos

a) Explorar e integrar fontes de dados de mídias sociais que serão usadas como recursos para mensurar o DES, considerando leis referentes a dados e políticas de privacidade de dados;
b) Definir parâmetros como forma de mensuração do escore de exposição digital;
c) Utilizar uma LLM capaz de identificar as informações sensíveis de cada usuário;
d) Desenvolver uma plataforma para demonstrar os principais marcadores identificados pelo modelo e mostrar o impacto geral em um conjunto de usuários da plataforma.

1.2. Justificativa

Os últimos anos confirmaram que a exposição maciça de dados pessoais é uma prática frequente. Em janeiro de 2021, o chamado "*vazamento do fim do mundo*" expôs informações de 223 milhões de brasileiros vivos e mortos, incluindo CPF, endereço, renda e vínculos bancários (CNN BRASIL, 2021). Em escala global, o incidente batizado de MOAB – *Mother of All Breaches*, revelado em janeiro de 2024, agregou mais de 26 bilhões de registros em um único repositório, confirmando que esses dados continuam em circulação, reutilização, o que potencializa fraudes ao redor do globo (CYBERNEWS, 2024).

Além da mera disponibilidade dessas informações, vetores técnicos historicamente negligenciados agravam o risco. O protocolo SS7, ainda usado na sinalização de redes móveis, permite a interceptação de SMS e chamadas. Em setembro de 2024 o canal Veritasium demonstrou publicamente como é trivial clonar mensagens de autenticação de dois fatores, enfraquecendo os próprios mecanismos de segurança implantados para conter vazamentos (VERITASIUM, 2024). Isso demonstrou como o usuário está suscetível a ataques sem que o usuário sequer saiba ou faça qualquer interação.

Paralelamente, surgem novas iniciativas de proteção, como as *Passkeys* (FIDO2/WebAuthn), cujo uso ultrapassou 15 bilhões de contas em dezembro de 2024, impulsionado pela adoção de empresas como Google, Amazon e Microsoft (FIDO ALLIANCE, 2024). Embora representem um avanço

significativo na prevenção de ataques de *phishing* e no fortalecimento da autenticação, essas soluções ainda não alcançam a totalidade dos usuários e tampouco mitigam os efeitos de informações vazadas por autoexposição digital.

Diante desse cenário, evidencia-se a ausência de uma métrica objetiva e acessível que permita ao cidadão compreender o grau de exposição da própria vida digital e, a partir disso, adotar medidas proporcionais de proteção. O projeto *Data Exposure Score* (DES) surge com o propósito de preencher essa lacuna, propondo uma forma de quantificação individual da exposição em redes sociais públicas.

O projeto DES pretende:

- Quantificar a exposição individual por meio de um escore similar ao de crédito, calculado a partir de bases públicas de redes sociais;
- Evidenciar o risco de superexposição digital nas mídias sociais;
- Incentivar boas práticas, oferecendo recomendações sobre os cuidados necessários ao publicar conteúdo *online.*

Dessa forma, o estudo aborda uma lacuna evidente na literatura e na prática: a ausência de um indicador comparável que oriente políticas públicas, programas educacionais e investimentos corporativos em cibersegurança.

1.3. Definição do escopo, contextualização e oportunidades

O sistema busca traduzir o comportamento de exposição em uma métrica quantitativa e interpretável, permitindo que os usuários compreendam o impacto de suas próprias postagens sobre sua privacidade e segurança.

O escopo do projeto concentra-se na análise de dados manifestamente públicos provenientes da rede social Bluesky, cuja API aberta possibilita a coleta ética e transparente das informações. A escolha dessa plataforma garante aderência à Lei Geral de Proteção de Dados (LGPD) e evita o uso de bases sensíveis ou vazadas, preservando a integridade jurídica e metodológica do estudo.

O DES não monitora indivíduos específicos nem realiza qualquer tipo de identificação pessoal. A abordagem é estatística e agregada, voltada à detecção de padrões coletivos de exposição e à criação de indicadores que auxiliem pesquisadores, educadores e profissionais de segurança da informação a compreender o comportamento digital contemporâneo.

Ao propor uma métrica padronizada para o risco de exposição, o projeto dialoga com a segurança da informação e a educação digital, destacando oportunidades de aplicação prática em campanhas de conscientização, políticas públicas e programas corporativos de cibersegurança. Espera-se que o DES possa contribuir para um debate mais amplo sobre privacidade e responsabilidade no ambiente online, oferecendo um instrumento mensurável que transforma o conceito abstrato de vulnerabilidade digital em dados objetivos, comparáveis e reproduzíveis.

2. Métodos e Tecnologias

O desenvolvimento do *Data Exposure Score* (DES) foi estruturado em etapas sequenciais e interdependentes, combinando fundamentos técnicos, legais e metodológicos para garantir a conformidade ética e a validade científica do sistema.

Foi conduzida uma análise das implicações da Lei Geral de Proteção de Dados (LGPD) (BRASIL, 2018), que orientou todas as decisões referentes à coleta, tratamento e armazenamento de informações. Essa avaliação assegurou que o projeto utilizasse apenas dados manifestamente públicos, respeitando os princípios de finalidade, necessidade e transparência previstos na legislação brasileira.

Com base nessas diretrizes, a rede social Bluesky foi definida como o ambiente primário de coleta de dados, em virtude de sua API aberta (AtProto) e política de acesso ético a conteúdos públicos. Essa escolha eliminou barreiras financeiras e legais presentes em outras plataformas e viabilizou a formação de um conjunto representativo de postagens e metadados.

O processamento das informações e a infraestrutura de armazenamento foram implementados com tecnologias em nuvem, priorizando escalabilidade e controle de custos. Os dados são manipulados por *scripts* em Python e armazenados no Amazon DocumentDB, um banco de dados orientado a documentos que oferecem flexibilidade para lidar com informações semiestruturadas.

A etapa de análise automatizada é conduzida por dois modelos: um modelo LLM executado via Amazon Bedrock e um modelo da OpenAI acessado por meio de uma API gerenciada.

O conjunto dessas tecnologias permite que o DES opere como um pipeline modular e escalável, no qual cada etapa — da coleta à inferência — pode ser aprimorada ou substituída conforme a evolução dos modelos e das ferramentas disponíveis, mantendo a integridade metodológica do sistema. A Figura 1 apresenta um esquema da interação das tecnologias utilizadas.

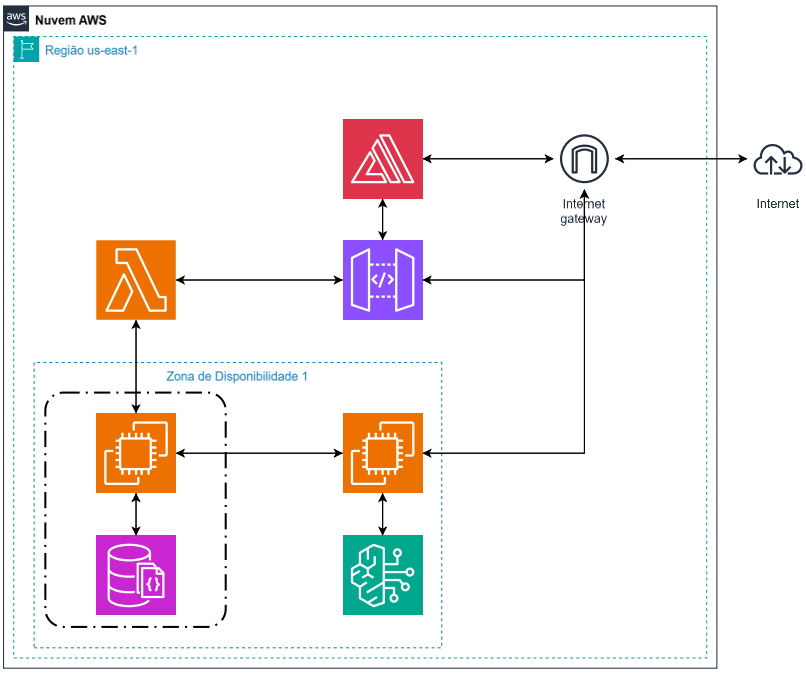


Figura 1 – Esquema da Interação das Tecnologias Utilizadas.

3. Revisão Bibliográfica

Apesar de muitos desses dados estarem fragmentados, anonimizados ou parcialmente ocultos, há um risco crescente de exposição de informações sensíveis. A possibilidade de cruzar diferentes conjuntos de dados permite a reidentificação de indivíduos, tornando vulneráveis informações que antes eram consideradas seguras. Essa preocupação é destacada por Mayer-Schönberger e Cukier (2013), que discutem a captura, a persistência indesejada e o uso secundário de dados — muitas vezes para finalidades completamente distintas daquelas para as quais foram originalmente coletados. Essa dinâmica amplia os riscos associados à análise massiva de informações, sobretudo quando combinada com dados de redes sociais e outras fontes públicas.

Além dos riscos técnicos, a exposição de dados é frequentemente facilitada por práticas de engenharia social. Mitnick (2002) define essa técnica como a arte de enganar alguém para obter informações ou acesso a algo de valor. Em sua obra *The Art of Deception*, o autor argumenta que o fator humano é o elo mais frágil na segurança da informação. Assim, ataques bem-sucedidos muitas vezes se iniciam com interações humanas que exploram gatilhos psicológicos como autoridade, urgência e simpatia. Um simples telefonema, mensagem direta ou e-mail malicioso pode ser suficiente para comprometer barreiras técnicas altamente sofisticadas.

A exposição digital de um indivíduo é resultado da interação entre a escala massiva de coleta de dados promovida por grandes empresas de tecnologia (*Big Techs*) e as vulnerabilidades humanas associadas ao comportamento online. Mesmo com o avanço de tecnologias de segurança — como criptografia, autenticação multifatorial e anonimização de dados —, o fator humano continua sendo determinante para a ocorrência de vazamentos. Uma única ação impensada, como clicar em um link de *phishing* ou compartilhar informações pessoais em ambientes públicos, pode comprometer toda uma infraestrutura de proteção.

Nesse contexto, o DES surge como uma abordagem metodológica voltada à mensuração da exposição digital com base em dados públicos de redes sociais. O sistema propõe um modelo de avaliação que permite quantificar o risco associado à superexposição de informações pessoais, contribuindo para o fortalecimento da cultura de segurança informacional e para a educação digital preventiva. Ao traduzir evidências empíricas em indicadores mensuráveis, o DES preenche uma lacuna existente entre as práticas de conscientização e a ausência de métricas objetivas de vulnerabilidade.

4. Desenvolvimento

O desenvolvimento do *Data Exposure Score* (DES) foi estruturado em uma arquitetura modular, composta por quatro etapas principais: coleta de dados, amostragem estatística, inferência automatizada e cálculo do escore de exposição. Essa abordagem garante escalabilidade, rastreabilidade e conformidade ética com as diretrizes da LGPD, permitindo que o sistema opere sobre dados manifestamente públicos de forma segura e reprodutível.

4.1. Coleta de dados da Rede Social Bluesky

O processo de coleta utilizou a rede social Bluesky como fonte primária de dados, escolhida por oferecer API aberta (AtProto) e política transparente de acesso a conteúdos públicos. Essa opção

eliminou barreiras de custo e viabilizou uma coleta ética, em conformidade com as normas de privacidade e uso de dados.

Foram desenvolvidos scripts em Python para extração automatizada de postagens públicas, metadados e informações de perfis com maior volume de interação. O sistema operou de modo distribuído, com múltiplas instâncias de execução paralela e controle de concorrência, bloqueio (locks) e verificação de duplicidade.

O resultado foi um conjunto de aproximadamente 8,19 GB de dados, correspondente a cerca de 92 milhões de publicações de mais de 500 mil usuários. Esses dados foram armazenados em Amazon DocumentDB (compatível com MongoDB), banco de dados orientado a documentos que oferece modo serverless e escalabilidade automática, garantindo desempenho consistente e eficiência de custos.

4.1.1. Foco em Redes Sociais

Diante das restrições legais e da inviabilidade técnica do uso de bases públicas genéricas, o escopo do projeto foi redefinido. A análise passou a se concentrar em um ambiente específico: as redes sociais. Esta escolha se justifica por serem plataformas onde os dados são, em grande parte, "manifestamente públicos pelo titular", conforme uma das hipóteses de tratamento de dados previstas pela LGPD, desde que respeitados os direitos e liberdades fundamentais do titular garantidos pela Constituição Federal.

Em seguida, a fase de seleção da plataforma foi analisada, considerando critérios técnicos e econômicos. Inicialmente, o caso de uso final do projeto seria direcionado à plataforma X (antigo Twitter). No entanto, ao conduzir uma análise de sua Interface de Programação de Aplicações (API), assim como a de outras redes consolidadas como o LinkedIn e Meta, revelou barreiras na disponibilidade das informações, além de custos proibitivos para a coleta de dados em larga escala. Diante dessas restrições, a rede social Bluesky foi selecionada como a fonte de dados primária. Por oferecer uma API aberta e gratuita, a plataforma removeu as barreiras financeiras e permitiu a extração do volume de dados necessário para o desenvolvimento e a validação do modelo proposto. Portanto, a estratégia metodológica consiste em utilizar os dados coletados como insumos para um modelo de LLM open source, utilizá-la de maneira massiva nos usuários da plataforma e construir relatórios de exposição digital nas principais segmentações de público (faixa etária, sexo, região).

4.1.2. Amostragem Estatística da base de dados

Devido ao grande volume de informações coletadas, foi necessário definir uma amostra representativa que viabilizasse a análise estatística sem comprometer os recursos computacionais.

O cálculo amostral considerou um nível de confiança de 95% e margem de erro de 0,01 na proporção, utilizando a fórmula:

$$n_0 = \frac{Z^2 \times p \times (1-p)}{E^2}$$

Sendo: Z = 1,96 (quantil da distribuição normal para 95% de confiança); p = 0,5 (pior caso para maximizar a variância populacional); E = 0,01 (margem de erro máxima tolerada).

O resultado indicou um tamanho mínimo de 9.604 usuários, que foi arredondado para 10.000 para assegurar robustez e simplicidade operacional. Essa amostra foi extraída de forma aleatória estratificada, equilibrando diversidade de perfis e consistência de padrões de interação. Assim, o conjunto amostral representa adequadamente o universo total de usuários, permitindo a inferência estatística do nível médio de exposição digital na plataforma Bluesky.

4.2. Geração do *Dataset* e do *Prompt*

Um dos principais desafios metodológicos identificados no projeto foi a inexistência de um conjunto de dados rotulado que pudesse servir de referência para validar as inferências do modelo de linguagem. A ausência de um resumo compilado, listando as informações exibidas pelos usuários na plataforma, exigiu a construção de um padrão próprio de anotação manual, elaborado a partir das postagens coletadas na etapa anterior.

Para contornar essa limitação, foi desenvolvida uma estrutura de dados em formato JSON, definindo os parâmetros esperados para a análise de exposição. Essa estrutura contém campos booleanos que indicam, respectivamente, a presença ou ausência de informações pessoais nas postagens públicas dos usuários. A definição desses campos foi construída com base em um mapeamento preliminar das categorias de exposição digital mais recorrentes, como menções a nome, localização, profissão, dados de contato ou preferências pessoais.

Cada registro do dataset corresponde a um usuário, contendo suas postagens públicas e o JSON correspondente à análise esperada. Esse formato padronizado permite que o modelo de linguagem *open source* receba o *prompt* de forma estruturada, garantindo consistência e reprodutibilidade nas inferências.

O prompt é composto por um conjunto de instruções descritivas que orientam o modelo a interpretar o conteúdo textual e gerar a saída no formato definido. As postagens dos usuários são inseridas nesse prompt, que é então processado pelo modelo de linguagem Llama 3.2 90B Instruct, resultando em um JSON final com os valores booleanos preenchidos.

O formato JSON adotado, como ilustrado na Figura 2, também facilita a integração direta com a base de dados no Amazon DocumentDB, permitindo o armazenamento e a recuperação estruturada dos resultados para análises subsequentes.

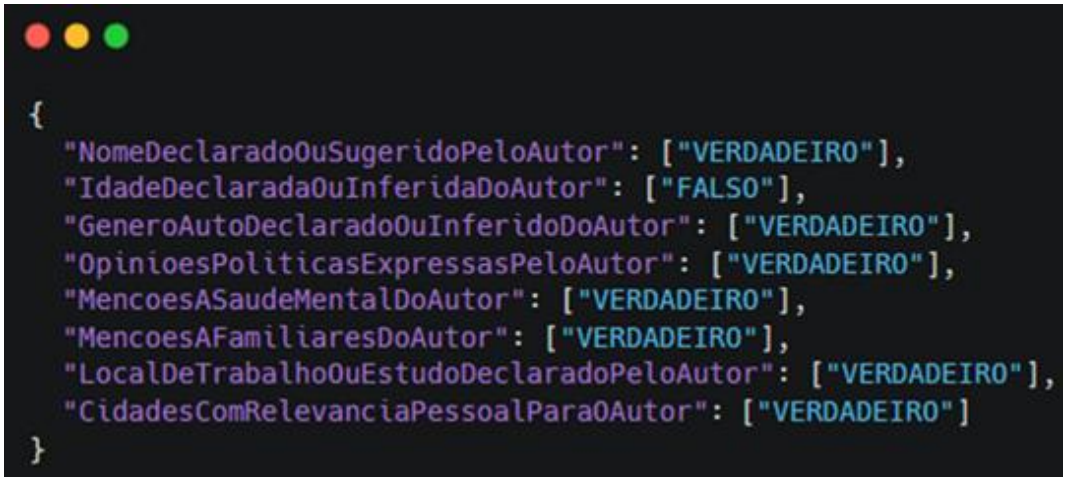


Figura 2 – Estrutura Simplificada de Output de dados para rotulagem de perfis.

4.3. Execução dos Modelos de IA

A etapa de execução do modelo de linguagem constitui o núcleo do sistema *Data Exposure Score* (DES), sendo responsável pela interpretação das postagens e pela geração dos resultados padronizados utilizados no cálculo do escore.

No projeto, o modelo é encarregado de analisar o conteúdo textual das postagens de cada usuário e gerar uma saída estruturada em formato JSON, contendo valores booleanos que indicam a presença ou ausência de informações sensíveis. Exemplos dessas categorias incluem “Contato pessoal”, “Informação financeira”, “Localização” e “Opinião sensível”, entre outras.

Para esta etapa, o sistema foi desenvolvido para integrar-se com múltiplas APIs de modelos gerenciados, especificamente o Amazon Bedrock e a API da OpenAI. Essa abordagem flexível permite ao projeto aproveitar os ganhos de escalabilidade, segurança e previsibilidade de custos oferecidos por ambas as plataformas, permitindo selecionar o *endpoint* mais adequado para cada cenário de análise.

O fluxo de execução atualizado segue três etapas principais, independentemente do provedor escolhido:

1. Montagem dos lotes de análise: As postagens são agrupadas em lotes (batches) compostos por um número fixo de usuários, de modo a otimizar o envio de requisições e o tempo de processamento.
2. Envio para a API do modelo: Cada lote é enviado ao endpoint de inferência apropriado. No caso do Amazon Bedrock, podem ser utilizados modelos como o Llama 3.2 90B Instruct; no caso da API da OpenAI, modelos da família GPT (como GPT-4o). O modelo selecionado processa o texto e interpreta o conteúdo conforme o formato definido.
3. Recebimento e armazenamento dos resultados: A API respectiva (seja Bedrock ou OpenAI) retorna um JSON padronizado com as chaves booleanas correspondentes às categorias de exposição identificadas. Esses resultados são então armazenados no Amazon DocumentDB, vinculados ao identificador de cada usuário analisado.

Essa integração garante consistência e rastreabilidade entre as etapas de coleta, processamento e armazenamento. Cada inferência executada corresponde a um usuário analisado, e os resultados alimentam diretamente o cálculo estatístico e ponderado do *Data Exposure Score* (DES).

4.4. Armazenamento e Flexibilidade do Amazon DocumentDB

O armazenamento dos dados foi implementado em Amazon DocumentDB (compatível com MongoDB), um banco de dados não relacional orientado a documentos. Essa escolha foi motivada por sua capacidade de lidar com informações semiestruturadas e não estruturadas — características comuns em postagens de redes sociais.

O DocumentDB oferece vantagens como o modo *serverless*, que permite o estado de hibernação quando não utilizado, e escalabilidade automática em momentos de alta demanda. Essa abordagem garante flexibilidade e eficiência de custos.

Além disso, o modelo de documentos possibilita a inclusão de novos atributos ao longo do tempo, sem necessidade de redefinir o *schema* do banco. Essa propriedade foi essencial para o projeto, que passou por diferentes estágios de coleta e teste, ajustando as variáveis utilizadas para o cálculo do escore DES. Essa flexibilidade estrutural também permitiu testar múltiplas versões do banco com diferentes tamanhos e tipos de dados, sem grandes impactos sobre a integridade ou a arquitetura geral do sistema.

4.5. Desenvolvimento do Modelo e Cálculo do Escore de Exposição

Com o conjunto de dados e a estrutura do *prompt* devidamente estabelecidos, esta seção descreve o processo de aplicação do modelo de linguagem de código aberto e a metodologia de cálculo do escore DES.

4.5.1. O *Data Exposure Score* (DES)

O *Data Exposure Score (DES)* quantifica o nível de exposição digital de um indivíduo a partir das informações sensíveis detectadas em suas publicações. O cálculo utiliza uma estrutura hierárquica inspirada no AHP (Analytic Hierarchy Process) em dois níveis, que separa: (i) a importância relativa dos critérios de avaliação e (ii) a relevância das categorias de dados dentro de cada critério. Abaixo segue um passo a passo, demonstrando como foi feito este cálculo:

1. Definição — critérios (ex.: *Impacto*, *Explorabilidade*, *Existência*) e categorias (informação financeira, documentos pessoais, localização, contato, rotina, afiliação, hobbies). A Tabela 2 apresenta valores de referência.
2. Comparações pareadas (nível 1) — construir a matriz entre os critérios e extrair o vetor de pesos $\mathrm{w} = (w_1, \ldots, w_m)$, normalizado tal que $\sum_c w_c = 1$.
3. Comparações pareadas (nível 2) — para cada critério c construir a matriz entre as n categorias e extrair o vetor de prioridades $\mathrm{p}_c = (p_{c,1}, \ldots, p_{c,n})$ com $\sum_j p_{c,j} = 1$.
4. Coerência — calcular o Consistency Ratio (CR) para cada matriz; se $CR > 0{,}10$, ajustar as comparações até $CR \leq 0{,}10$.

5. Peso global por categoria — combinar os níveis:

$$W_j = \sum_{c=1}^{m} w_c \cdot p_{c,j}, j = 1, \dots, n,$$

os W_j resultantes são normalizados por construção.

6. Variável de exposição — para cada usuário e categoria $j$define-se V_j. Neste estudo usa-se a forma binária $V_j \in \{0,1\}$(1 = detectada no JSON; 0 = não detectada).
7. 7. Cálculo do DES — somatório ponderado

$$S = \sum_{j=1}^{n} W_j \cdot V_j, S \in [0,1],$$

e escala final invertida:

Assim, $DES = 1000$ indica ausência de exposição detectada; $DES = 0$exposição máxima segundo os critérios adotados.

A **Tabela 2** abaixo fornece os valores referenciais de *Impacto* e *Explorabilidade* por categoria. Esses valores são insumo para obter os vetores $p_c$do nível 2, pela seguinte maneira:

- Conversão para AHP (recomendado para verificação de coerência): para cada critério, converte-se os valores $v_i$da Tabela 2 em uma matriz pareada por razões $a_{ij} = v_i/v_j$; extrai-se o autovetor p_c e calcula-se o CR (ajustar se $CR > 0{,}10$).

Em ambos os casos, combinar os $p_c$com os pesos de nível 1 (w_c) para obter W_j.

Tabela 2: Estrutura de Ponderação de Parâmetros para o Cálculo do DES (Modelo Expandido)

Categoria	Descrição	Impacto (I)	Explorabilidade	Justificativa
Informação Financeira	Menção a salários, bancos, cartões	10	8	Risco direto de fraude financeira e engenharia social direcionada.
Documentos Pessoais	Menção a CPF, RG, CNH	10	7	Risco crítico de roubo de identidade, embora a exploração exija mais passos.
Localização em Tempo Real	Check-ins, menções a "estou em..."	8	9	Risco iminente de segurança física (stalking, roubo). Altamente explorável.
Contato Pessoal	E-mail, número de telefone	8	10	Vetor direto e de baixa complexidade para phishing, smishing, e

				contato indesejado.
Rotina/Hábitos	Horários de trabalho, locais frequentados	6	6	Permite traçar perfil de comportamento para ataques de engenharia social e física.
Afiliação Política/Religiosa	Declarações de posicionamento	4	5	Risco de discriminação, assédio direcionado, e perfilamento ideológico.
Hobbies e Interesses	Gostos, atividades de lazer	2	4	Baixo risco direto, mas útil para criar pretextos críveis em engenharia social.

Consideradas as mais sensíveis, as informações financeiras implicam risco econômico direto (fraude, transações não autorizadas) e, quando combinadas com identificadores, reduzem muito o esforço necessário para exploração — justificando peso próximo aos identificadores (XIA et al., 2023).

De modo semelhante, identificadores pessoais têm alto potencial de dano, pois permitem associação direta à pessoa e viabilizam fraudes e usurpação de identidade; por isso, uma exposição mesmo isolada eleva substancialmente o risco medido pelo DES.

Números de telefone e outros tipos de contato atuam como vetores para engenharia social (*phishing*) e ampliam o alcance do dano; isoladamente, apresentam severidade média, mas sua combinação com outros atributos aumenta a probabilidade de ataque, recebendo peso intermediário.

Dados sensíveis possuem elevada severidade ética e jurídica: a LGPD impõe tratamento diferenciado por seu potencial discriminatório; mesmo que a probabilidade de reidentificação varie, o impacto social e legal de sua divulgação justifica peso significativo no DES (BRASIL, 2018; ANPD, 2023).

Da mesma forma, dados de localização favorecem reidentificação por correlação temporal e espacial (por exemplo, residência e local de trabalho) e podem implicar riscos físicos como *stalking*; pela capacidade de *linkability* e inferência de rotinas, merecem peso relevante no escore.

4.6. Dashboards de Acompanhamento

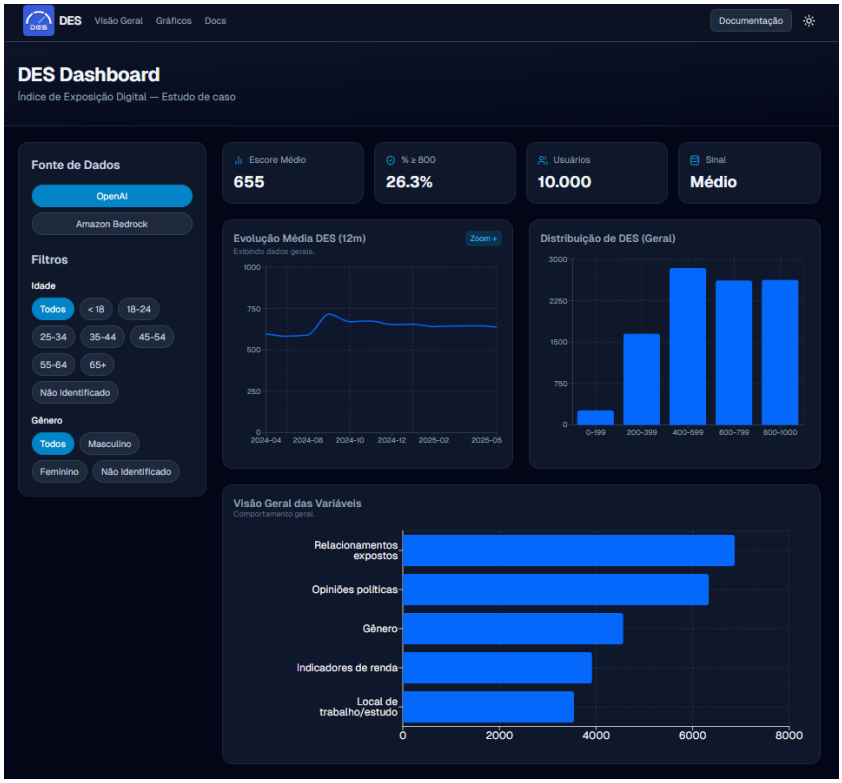
O projeto foi concebido com uma visão de evolução contínua, voltada à ampliação de sua aplicabilidade e impacto social. Para materializar essa estratégia, foi implementado um dashboard de acompanhamento desenvolvido em Next.js, que serve como a interface central de inteligência do sistema, traduzindo o complexo cálculo do Data Exposure Score (DES) em visualizações intuitivas e acionáveis.

A plataforma oferece uma visão holística da exposição digital, permitindo o monitoramento dinâmico de grandes volumes de dados. Através de uma arquitetura flexível, o sistema integra diferentes fontes de processamento de IA (como modelos da OpenAI e Amazon Bedrock), garantindo robustez e escalabilidade na análise.

O painel foi desenhado para oferecer diagnósticos em múltiplas camadas:

- Segmentação e Filtragem: O sistema permite recortes demográficos precisos — por faixas etárias e gênero — possibilitando identificar quais grupos estão mais vulneráveis.
- Análise Temporal e de Tendências: Gráficos de evolução histórica permitem acompanhar a flutuação do risco ao longo do tempo, validando a eficácia de medidas de segurança adotadas.
- Mapeamento de Vulnerabilidades: Mais do que apenas apontar um "score" geral, o dashboard detalha as variáveis específicas de risco, destacando quais aspectos da vida do usuário (como relacionamentos, opiniões políticas ou dados profissionais) estão mais expostos publicamente.

Essa estrutura não apenas demonstra o comportamento técnico do DES, mas fundamenta as orientações contextuais oferecidas ao usuário. Ao visualizar exatamente onde reside o perigo, a ferramenta deixa de ser apenas um medidor passivo e torna-se um instrumento ativo de educação digital, facilitando a mudança de hábitos e o fortalecimento da privacidade online.



5. Resultados

5.1. Rotulagem Automatizada

O dataset de validação utilizado neste estudo foi construído a partir de postagens públicas da rede social Bluesky, coletadas via API aberta AtProto, no período de 24/05/2025 a 29/05/2025. A coleta bruta

resultou em aproximadamente 92 milhões de publicações de mais de 500 mil usuários (cerca de 8,19 GB de dados), armazenadas em um banco Amazon DocumentDB compatível com MongoDB, conforme descrito na Seção 4.1.

A partir desse universo, foi definida uma amostra estatística de 10.000 usuários, calculada com 95% de confiança e margem de erro máxima de 1% na proporção, que serviu como base para os experimentos com o Data Exposure Score (DES). Para compor o dataset de rotulagem, foram consideradas todas as postagens públicas desses usuários dentro da janela de coleta, totalizando 1,4 milhões de postagens únicas, distribuídas entre os usuários e organizadas em uma estrutura própria de anotação.

O sistema analisa as amostras de postagens e as classifica em duas categorias principais:

- VERDADEIRO: quando há identificação de informação sensível pertencente ao próprio autor do conteúdo (por exemplo, dados de contato, localização, rotina, informações financeiras ou documentos pessoais);
- FALSO: quando não é identificada informação sensível relevante que contribua para o aumento do risco de exposição digital do autor.

Cada registro do dataset é representado por um objeto em estrutura JSON padronizada, que consolida o texto analisado e um conjunto de campos booleanos para as dimensões de exposição avaliadas (como informação financeira, documentos pessoais, localização em tempo real, contato pessoal, rotina/hábitos, afiliações e hobbies/interesses). Essa estrutura JSON é utilizada como entrada para o modelo de linguagem open source e como formato de saída esperado: o modelo interpreta o prompt com as postagens do usuário e retorna o JSON preenchido com valores verdade/falso para cada categoria, possibilitando a comparação direta com os rótulos de referência e o cálculo das métricas de desempenho na validação.

5.2. Uso de LLM para Identificação de Exposição

A precisão do *Data Exposure Score* (DES) depende diretamente da capacidade do modelo de linguagem (LLM) em analisar o conteúdo das postagens e identificar informações sensíveis. O projeto utilizou os modelos GPT-4o da OpenAI e o modelo Llama 3.2 90B Instruct executado via Amazon Bedrock para interpretar o texto e preencher a estrutura JSON padronizada com valores booleanos.

- Modelo "Llama 3.2 90B Instruct": Ao analisar os 10.000 usuários, este modelo resultou em um Escore Médio de 893 ("Alto"). Isso sugere que o modelo falhou em detectar muitas informações sensíveis, gerando uma falsa sensação de segurança. Isso foi concluído devido a dois principais fatores: em primeiro lugar, a diferença quantitativa de categorização e identificação (por idade e gênero) entre os dois, e em segundo lugar, a diferença significativa do score que cada um dos modelos trouxe, considerando que a amostra era exatamente a mesma. Ambos serão tratados a seguir.

- Com relação ao primeiro aspecto, foi possível notar uma grande discrepância entre o comportamento dos dois modelos: o modelo da OpenAI não conseguiu reconhecer classificações de gênero e idade de 5.236 usuários (aproximadamente 52% da amostra), enquanto o Llama deixou de identificar (gênero e idade, juntos) 8.666 usuários (cerca de 86% da amostra). Dessa forma, verifica-se uma diferença de aproximadamente 34 pontos percentuais entre os modelos, representando uma proporção maior de usuários que o Llama não conseguiu classificar em comparação ao modelo da OpenAI.

- Como segundo ponto, o modelo da OpenAI: Utilizando o modelo do projeto, o Escore Médio caiu para 655 ("Médio"). Isso indica que o Llama 3.2 90B Instruct foi significativamente menos eficaz em identificar exposições (como ‘Relacionamentos expostos’ e ‘Indicadores de renda’), que foram detectadas pelo modelo da OpenAI. A análise comparativa revelou uma divergência crítica na eficácia de detecção. Enquanto o modelo Llama 3.2 90B Instruct atribuiu um Escore Médio de 893 (Nível 'Alto/Seguro'), o modelo da OpenAI reclassificou a mesma amostra para 655 (Nível 'Médio').

Essas duas discrepâncias evidenciam que o score elevado do Llama não reflete segurança real, mas sim uma limitação técnica do modelo em realizar inferências semânticas complexas. O modelo falhou sistematicamente em detectar exposições indiretas (especificamente nas categorias de 'Relacionamentos' e 'Indicadores de Renda') gerando uma taxa elevada de falsos negativos. Em contraste, o modelo da OpenAI demonstrou maior sensibilidade contextual, identificando corretamente padrões de linguagem que denotam vulnerabilidade, resultando em uma métrica de risco mais rigorosa e fidedigna à realidade dos dados expostos. A partir disso, faz-se necessário ressaltar que um LLM robusto e bem instruído muda os resultados de forma crucial. Modelos menos potentes podem falhar em interpretar o contexto e as nuances das postagens, subestimando o real nível de exposição digital do usuário.

5.3. Eficácia da Metodologia AHP no cálculo do Escore

Para que o DES refletisse o risco real, não bastava apenas contar as exposições; era preciso ponderá-las qualitativamente. O estudo utilizou uma estrutura hierárquica baseada no AHP (Analytic Hierarchy Process) para calcular o escore final. A eficácia dessa metodologia foi validada pela consistência matemática dos pesos atribuídos, o que permitiu transformar julgamentos subjetivos em uma escala objetiva de risco, priorizando critérios estruturais como "Impacto" e "Explorabilidade".

Essa calibração fica evidente ao analisar a Tabela 2:

- Informação Financeira: Classificada com "Impacto" 10 e "Explorabilidade" 8.
- Hobbies e Interesses: Classificados com "Impacto" 2 e "Explorabilidade" 4.

Ao aplicar essa matriz de pesos, o algoritmo do AHP assegura matematicamente que a exposição de dados financeiros penalize o escore de forma muito mais severa do que a menção a um hobby, alinhando o cálculo à realidade da segurança da informação. Além disso, a estrutura do modelo atribuiu uma preponderância global ao critério "Impacto", garantindo que a severidade do dano potencial seja o fator determinante no cálculo final. Isso assegura que o DES (na escala de 0 a 1000) não seja apenas um contador, mas um quantificador preciso da gravidade da vulnerabilidade.

5.4. Replicabilidade do Cálculo do DES em outras plataformas

O estudo confirma que o Data Exposure Score (DES) foi desenvolvido como um modelo replicável, podendo ser adaptado para analisar outros contextos de redes sociais além da Bluesky. A chave para essa replicabilidade é a arquitetura modular do sistema, que separa a coleta de dados da análise. O pilar dessa arquitetura é a estrutura JSON padronizada (vista na Figura 2 do documento).

O processo funciona da seguinte forma:

1. Coleta: Os dados são extraídos de uma plataforma (no estudo, a Bluesky).
2. Análise (Padronizada): O LLM recebe as postagens e é instruído (via *prompt*) a preencher o JSON padronizado, identificando categorias específicas (ex: "OpinioesPoliticas", "LocalDeTrabalho").
3. Cálculo: O escore DES é calculado usando a metodologia AHP sobre a saída JSON.

Desde que a coleta de dados de outra rede social (como X ou LinkedIn) seja viável e que os dados textuais possam ser alimentados no LLM, o mesmo prompt e a estrutura JSON podem ser usados para gerar o escore. Isso torna a metodologia DES uma ferramenta de análise de risco portátil e escalável.

5.5. Aplicabilidade do DES

O *Data Exposure Score* (DES) foi concebido para funcionar como um indicador de exposição digital, capaz de traduzir o comportamento de compartilhamento de informações em um valor numérico interpretável. Sua principal aplicabilidade está em quantificar e comparar níveis de exposição entre diferentes grupos de usuários, possibilitando análises mais objetivas sobre como os padrões de interação variam conforme o tipo de conteúdo publicado.

O sistema também se mostra relevante como ferramenta de conscientização, permitindo que indivíduos, instituições e pesquisadores visualizem de maneira clara o impacto da divulgação de informações pessoais em ambientes públicos. Ao associar o escore de exposição às categorias de dados detectadas, o DES fornece uma dimensão prática do risco informacional, reforçando o entendimento sobre a importância da privacidade digital.

Além disso, o indicador oferece potencial de aplicação em programas de educação digital, segurança da informação e políticas públicas, apoiando o desenvolvimento de estratégias voltadas à redução da exposição indevida de dados pessoais. Dessa forma, o DES se consolida como um instrumento integrador entre tecnologia, ética e comportamento digital, contribuindo para o avanço das discussões sobre privacidade e segurança nas redes sociais contemporâneas.

6. Conclusões

Este trabalho apresentou o desenvolvimento do sistema *Data Exposure Score* (DES) – uma solução inovadora para mensurar o grau de exposição digital de usuários em redes sociais.

A partir da análise de dados manifestamente públicos extraídos da plataforma Bluesky, foi possível criar um *pipeline* automatizado que identifica, categoriza e quantifica informações pessoais. Para a atribuição dos pesos de risco, o projeto fundamentou-se na metodologia AHP (*Analytic Hierarchy Process*), garantindo consistência matemática e decisória na hierarquização da sensibilidade dos dados expostos.

Os resultados apontam que a exposição digital é frequente e muitas vezes inconsciente: práticas cotidianas como menções a localização, rotinas e dados identificáveis tornam os próprios usuários vetores de vulnerabilidade, facilitando ataques de engenharia social e *phishing*.

O projeto enfrentou limitações jurídicas e financeiras: as restrições impostas pela LGPD impediram o uso de bases de vazamentos, e as barreiras de acesso às APIs de redes consolidadas limitaram a escalabilidade do sistema. Ainda assim, a estratégia de utilizar a Bluesky como fonte de dados e o Amazon DocumentDB como repositório mostrou-se eficaz para demonstrar a aplicabilidade e flexibilidade da arquitetura.

Como visão de longo prazo, o grupo prioriza a consolidação de um ecossistema 100% *open source* e a evolução do motor de cálculo. Pretende-se transitar da estrutura estática do AHP para o desenvolvimento de uma métrica proprietária e dinâmica, capaz de recalibrar os pesos automaticamente com base em novos padrões de ameaças digitais.

Conclui-se que o sistema DES preenche uma lacuna crítica na área de segurança da informação ao propor uma métrica acessível, educativa e escalável para avaliar o risco de exposição digital. A continuidade do projeto pode contribuir para políticas públicas, programas de capacitação digital e estratégias corporativas de cibersegurança, promovendo uma internet mais segura, consciente e democratizada.

Referências

AT&T TECH CHANNEL. *The step-by-step switch* [vídeo]. YouTube, 2013. Disponível em: https://youtu.be/xZePwin92cI?si=e1bVP0LOIxsy221C. Acesso em: 28 abr. 2025.

AUTORIDADE NACIONAL DE PROTEÇÃO DE DADOS — ANPD. *Estudo técnico sobre anonimização de dados na LGPD: uma visão de processo baseado em risco e técnicas computacionais*. Brasília: ANPD, 2023.

BLACK, C. et al. Ghost in the network. *Lighthouse Reports*, 2023. Disponível em: https://www.lighthousereports.com/investigation/ghost-in-the-network. Acesso em: 28 abr. 2025.

BRASIL. Lei nº 13.709, de 14 de agosto de 2018. Dispõe sobre o tratamento de dados pessoais e altera a Lei nº 12.965, de 23 de abril de 2014 (Marco Civil da Internet). *Diário Oficial da União*,

CISA. Avoiding Social Engineering and Phishing Attacks. Disponível em: https://www.cisa.gov/news-events/news/avoiding-social-engineering-and-phishing-attacks. Acesso em: 28 abr. 2025.

CNN BRASIL. Fotos e até salários estão entre os dados vazados de 223 milhões de brasileiros. 27 jan. 2021. Disponível em: https://www.cnnbrasil.com.br/tecnologia/fotos-e-ate-salarios-estao-entre-os-dados-vazados-de-223-milhoes-de-brasileiros/. Acesso em: 28 abr. 2025.

CYBERNEWS. Mother of all breaches reveals 26 billion records: what we know so far. 29 jan. 2024. Disponível em: https://cybernews.com/security/billions-passwords-credentials-leaked-mother-of-all-breaches/. Acesso em: 28 abr. 2025.

DICIO. Cibercultura. *Dicionário Online de Português*, 06 jul. 2017. Disponível em: https://www.dicio.com.br/cibercultura/. Acesso em: 09 jun. 2025.

ENGEL, T. SS7: locate. track. manipulate. [apresentação]. *31C3 – Chaos Communication Congress*, Hamburgo, 2014. Disponível em: https://youtu.be/-wu_pO5Z7Pk?si=xYhtwEuc_Qaqqfla. Acesso em: 28 abr. 2025.

FIDO ALLIANCE. Biometric update – Passkeys build momentum, enabling access to 15 billion online accounts. 16 dez. 2024. Disponível em: https://fidoalliance.org/biometric-update-passkeys-build-momentum-enabling-access-to-15-billion-online-accounts/. Acesso em: 28 abr. 2025.

IBM. Engenharia social: o que é e como se proteger. Disponível em: https://www.ibm.com/br-pt/topics/social-engineering. Acesso em: 28 abr. 2025.

MAYER-SCHÖNBERGER, V.; CUKIER, K. *Big data: a revolution that will transform how we live, work, and think*. Boston: Eamon Dolan/Houghton Mifflin Harcourt, 2013.

MITNICK, K.; SIMON, W. L. *The Art of Deception: Controlling the Human Element of Security*. Hoboken: Wiley, 2003.

NARAYANAN, A.; SHMATIKOV, V. Robust de-anonymization of large sparse datasets. In: *Proceedings of the 2008 IEEE Symposium on Security and Privacy (SP)*, 2008.

NOHL, K. Mobile self-defense. [apresentação]. *31C3 – Chaos Communication Congress*, 2014. Disponível em: https://youtu.be/nRdJ0vaQt0o?si=Bojv_Qq9lXasqQA3. Acesso em: 28 abr. 2025.

SWEENEY, L. k-Anonymity: a model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems*, v. 10, n. 5, p. 557–570, 2002.

THE ECONOMIST. It is dangerously easy to hack the world's phones. 2024. Disponível em: https://www.economist.com/science-and-technology/2024/05/17/it-is-dangerously-easy-to-hack-the-worlds-phones. Acesso em: 28 abr. 2025.

VERITASIUM. Exposing the flaw in our phone system [vídeo]. YouTube, 21 set. 2024. Disponível em: https://www.youtube.com/watch?v=wVyu7NB7W6Y. Acesso em: 28 abr. 2025.

XIA, W. et al. Managing re-identification risks while providing access to microdata: a practical framework. *Journal of the American Medical Informatics Association*, 2023.

ZUBOFF, Shoshana. *A era do capitalismo de vigilância: a luta por um futuro humano na nova fronteira do poder*. Rio de Janeiro: Intrínseca, 2021.