VeritaPlugin: Uma Extensão de Navegador para Detecção Semântica de Fraudes no Facebook

Bruna Assis Norões[1], Bianca Monsores da Silva Rocha[1], Rebeca Motta[1],

[1]Instituto de Computação - Universidade Federal Fluminense (UFF)
Av. Gal. Milton Tavares de Souza - São Domingos - Niterói - RJ

{brunanoroes,biancamsr,rcmotta}@id.uff.br

Abstract. *Social engineering on social networks exploits vulnerabilities to deceive users, rendering traditional technical defenses insufficient. This work presents VeritaPlugin, a browser extension that detects fraud on Facebook through a hybrid pipeline using BERTimbau, deterministic RAG, and GPT-4o. The architecture operates in compliance with the LGPD (Brazilian General Data Protection Law), and the result presents the user with the scam category, legal framework, and recommended actions. For calibration, the BrScamsFacebook dataset was built and made available, containing 450 instances of real scams from the Brazilian context. In the technical evaluation, the classifier obtained an F1-macro of 0.763 ± 0.034 in the k=5 cross-validation, exceeding the baseline.*

Resumo. *A Engenharia Social em redes sociais explora vulnerabilidades para iludir usuários, tornando defesas técnicas tradicionais insuficientes. Este trabalho apresenta o VeritaPlugin, uma extensão de navegador que detecta fraudes no Facebook por meio de um pipeline híbrido BERTimbau, RAG determinístico e GPT-4o. A arquitetura opera em conformidade com a LGPD e o resultado apresenta ao usuário a categoria do golpe, enquadramento legal e ações recomendadas. Para calibração, foi construído e disponibilizado o dataset BrScamsFacebook, com 450 instâncias de golpes reais do contexto brasileiro. Na avaliação técnica, o classificador obteve F1-macro de 0,763 ± 0,034 na validação cruzada k=5, superando o baseline.*

1. Introdução

Atualmente, as redes sociais fazem parte da vida diária tanto da sociedade brasileira quanto do restante do mundo. Estima-se que, em 2025, o Brasil contava com 144 milhões de usuários de redes sociais, correspondendo a 67,8% da população [Kemp 2025].

Através das redes, a comunicação é facilitada, ganhando novos patamares ao ser direta, instantânea e assíncrona, ocorrendo tão rápida quanto a velocidade da infraestrutura de internet disponível. Assim, a digitalização da comunicação transformou profundamente nossas interações, tornando redes como o Facebook [1], infraestruturas críticas do cotidiano com impacto social e econômico[Laor 2022].

Contudo, essas redes, que convertem o desejo humano de conexão em engrenagens algorítmicas, são exploradas por agentes maliciosos por meio da engenharia social [Silva et al. 2024]. Golpes em redes sociais atuam como uma ramificação específica de

[1]*https://www.facebook.com/facebook/about/*

ataque cibernético que, em vez de focar exclusivamente em falhas de software ou hardware, explora as vulnerabilidades da interação humana e o controle limitado de conteúdo presente nessas plataformas. Dessa forma, esses golpes focam em vulnerabilidades emocionais e comportamentais, tornando defesas puramente técnicas, como firewalls e filtros de spam, insuficientes [Daim et al. 2024]. Assim, o fator humano se torna um elo frágil na cibersegurança, sendo responsável pela maioria dos incidentes [Jamil et al. 2018].

Em casos em que ameaças nas redes sociais ocorrem, um agravante é a ausência de fontes neutras de validação. Por isso muitas vítimas não buscam ajuda externa por vergonha ou privacidade, mantendo-se sob a narrativa do golpista [Bleiman et al. 2025]. Além disso, soluções comerciais como antivírus e verificadores de URL têm limitações ao lidar com golpes em texto puro (comuns no Marketplace e Messenger), que não utilizam links externos maliciosos. Existe, portanto, uma lacuna para ferramentas que atuem na camada semântica, analisando a intenção do discurso em conformidade com a Lei Geral de Proteção de Dados Pessoais (LGPD) [BRASIL 2018].

Para endereçar essa lacuna, apresentamos o **VeritaPlugin**, uma extensão de navegador baseada em Inteligência Artificial (IA) que analisa textos no Facebook e identifica golpes de Engenharia Social, fornecendo ao usuário a categoria da ameaça, justificativas explicativas e enquadramento legal.

A questão central que orienta a pesquisa é: *Como uma solução em plugin seria capaz de classificar corretamente golpes digitais em textos do Facebook e gerar justificativas compreensíveis para usuários com diferentes níveis de letramento digital?* Para responder essa questão, a pesquisa realizada resultou no desenvolvimento do VeritaPlugin e trouxe outras contribuições:

- Disponibilização do VeritaPlugin: A criação de uma extensão de navegador para o Google Chrome, desenvolvida sob as especificações do Manifest V3, que atua diretamente na camada de apresentação (DOM) das páginas do Facebook.
- Pipeline Híbrido: A implementação de uma arquitetura que integra o modelo BERTimbau para classificação semântica em português; um mecanismo de RAG (*Retrieval-Augmented Generation*) determinístico, que recupera blocos de conhecimento jurídico e *modus operandi* específicos para a categoria de golpe detectada; e a geração de linguagem natural via GPT-4o para fornecer justificativas explicativas.
- Dataset BrScamsFacebook: A construção e disponibilização de um dataset representativo do contexto brasileiro de fraudes, composto por 450 instâncias balanceadas em seis categorias.
- Apoio ao Letramento Digital: Ao fornecer enquadramento legal detalhado e explicar os gatilhos de engenharia social detectados, a ferramenta busca promover o desenvolvimento do senso crítico dos usuários e ajudar a reconhecer padrões de manipulação de forma autônoma em interações futuras.

O restante deste artigo está organizado como segue: a seção 2 fundamenta os conceitos de engenharia social, taxonomia de golpes e arcabouço legal. A seção 3 detalha a metodologia, incluindo personas e requisitos. A seção 4 descreve a implementação do VeritaPlugin, o dataset BrScamsFacebook e a arquitetura do pipeline híbrido. A seção 5 apresenta o VeritaPlugin e discute os resultados da avaliação técnica. Por fim, a seção 6 apresenta conclusões e direções para trabalhos futuros.

2. Fundamentação Teórica

Esta seção apresenta a fundamentação teórica e o estado da arte necessários para compreender a arquitetura e o propósito do VeritaPlugin, focando na psicologia da persuasão, na classificação de fraudes no Brasil e na evolução das técnicas de detecção automatizada.

2.1. Engenharia Social

É definido como **ameaça cibernética** qualquer circunstância ou evento com potencial para impactar negativamente as operações organizacionais (incluindo missão, funções, imagem ou reputação), os ativos da organização ou indivíduos por meio de um sistema de informação [NIST 2021].

A engenharia social explora essas ameaças, mas se manifesta através da manipulação psicológica e o uso de princípios de persuasão para convencer indivíduos a realizarem ações ou revelarem informações que, de outra forma, não fariam [Mitnick and Simon 2003]. Diferente de vulnerabilidades de software, essa prática explora o comportamento humano, tornando difícil de mitigar apenas com defesas técnicas. Agentes maliciosos, nesse contexto, não buscam "hackear" sistemas, mas sim pessoas.

Para fundamentar essa análise, [Ferreira et al. 2015] compilaram uma estrutura unificada de princípios de persuasão voltada ao mundo online, realizada a partir de uma taxonomia de influência [Cialdini 2007]. Esses golpes exploram os princípios de persuasão: **autoridade** (simular entidades oficiais), **prova social** (mimetizar a maioria), **afinidade e decepção** (identidades falsas), **comprometimento e reciprocidade** (coerência com ações passadas) e **distração** (uso de urgência e medo).

Compreender os mecanismos psicológicos da Engenharia Social é o ponto de partida para a detecção automatizada de fraudes. Mas para que um sistema possa classificar e explicar ameaças de forma útil ao usuário, é necessário também mapear como esses mecanismos se manifestam em padrões concretos de ataque. A seção seguinte apresenta a taxonomia de golpes adotada neste trabalho, organizando as principais modalidades de fraude observadas no contexto brasileiro do Facebook.

2.2. Taxonomia de Golpes no Contexto Brasileiro

As táticas fraudulentas variam em método e objetivo, por isso, compreendê-las e categorizá-las é estratégico no desenvolvimento de respostas eficazes. Embora esses golpes explorem a natureza psicológica, os mecanismos de realização são diversos. Alguns exploram a construção e o abuso da identidade, outros se concentram na entrega de conteúdo malicioso. Isso se torna ainda mais preocupante se aliado a estratégias de escalabilidade na exploração da confiança dos usuários, como o uso de *bots* [Ariza et al. 2022].

A partir do trabalho de [Frasão et al. 2025], considerando os tipos de golpes mais frequentemente relacionados ao Facebook [Kotzias et al. 2025], este trabalho identifica golpes distintos organizando-os em cinco categorias (Figura 1): Golpes de Ganho Financeiro Ilusório, Golpes de Desinformação Digital, Fraudes em Lojas Virtuais Falsas, Ataques de Phishing e Roubo de Dados e Golpes Baseados em Relacionamento Romântico.

É importante ressaltar que no ecossistema digital, esses ataques podem operar de maneira conjunta. Fraudes digitais podem constituir campanhas heterogêneas que utilizam múltiplas táticas simultaneamente para potencializar sua eficácia e alcance. Dessa

Categoria	Definição
Golpes de Ganho Financeiro Ilusório	Compreende esquemas fraudulentos desenhados para induzir a vítima a realizar pagamentos antecipados ou ceder informações sensíveis sob a falsa promessa de uma grande recompensa futura (como prêmios, heranças ou retornos de investimento). Para isso, os é empregado técnicas de Engenharia Social que exploram a ingenuidade, a necessidade econômica ou a ganância do usuário.
Golpes de Desinformação Digital	Consiste na criação e disseminação massiva e deliberada de informações falsas ou manipuladas, frequentemente impulsionadas pelo uso de bots e contas falsas. O objetivo central é obter vantagens financeiras ou influenciar a opinião pública em larga escala, explorando notícias enviesadas e gatilhos emocionais da audiência, como medo, raiva e o senso de urgência.
Fraudes em Lojas Virtuais Falsas	Refere-se ao uso malicioso da infraestrutura de vendas das redes sociais (como o Marketplace) para atrair consumidores com ofertas irreais, clonando muitas vezes a identidade visual de marcas legítimas. A finalidade do golpe é receber o pagamento (sem a entrega do produto) ou capturar dados bancários e informações financeiras durante o falso processo de compra.
Ataques de Phishing e Roubo de Dados	Engloba campanhas focadas na extração sistemática de dados sensíveis, credenciais de acesso e informações financeiras para fins ilícitos. Por meio da Engenharia Social, os estelionatários mimetizam instituições legítimas (como bancos ou redes sociais) ou utilizam perfis falsos atraentes para disparar links maliciosos, comprometer dispositivos ou manipular a vítima a entregar suas credenciais voluntariamente.
Golpes Baseados em Relacionamento Romântico	Fundamenta-se na manipulação psicológica e na Engenharia Social para explorar a vulnerabilidade emocional, a empatia e a confiança da vítima. Através da criação de identidades falsas altamente atraentes, o estelionatário estabelece um vínculo afetivo intenso e acelerado. Uma vez consolidada essa relação íntima, o criminoso explora a vítima de duas formas principais: fabricando falsas crises ou tragédias urgentes para obter assistência financeira, ou induzindo o compartilhamento de mídias íntimas para posterior chantagem, exigindo dinheiro sob a ameaça de expor o conteúdo à rede de contatos da vítima.

Figura 1. Categoria de Golpes.

forma, a taxonomia dos crimes cibernéticos revela-se cada vez mais fluída. Portanto, compreender que há uma intersecção e a categorização múltipla dos ataques é importante para a mitigação de riscos e para a responsabilização legal dos agentes maliciosos.

Deste modo, é necessário distinguir os vetores de ataque de seus objetivos finais, que comumente se sobrepõem na execução do crime. O *Phishing* é frequentemente usado com outros golpes, como por exemplo com o *Identity Theft*. Esses dois golpes ocupam papéis complementares: o *phishing* é o meio (o link falso ou a mensagem enganosa) enquanto o roubo de dados é o fim (a extração de informações pessoais).

2.3. Trabalhos Relacionados

Para organização do trabalho, revisamos a literatura de detecção automática de fraudes em texto em três frentes: abordagens baseadas em *Machine Learning*(ML) clássico, abordagens baseadas em LLMs, e soluções centradas no letramento do usuário.

ML clássico e detecção binária. A primeira geração de soluções emprega vetorização por TF-IDF combinada a algoritmos como SVM, XGBoost e LightGBM para classificação binária fraude/não-fraude. [Nascimento et al. 2023] apresentam os datasets FraudWhatsApp.Br e FraudTelegram.Br com mensagens em português brasileiro e alcançam F1-score de 0,99 combinando TF-IDF, Bag of Words e nove classificadores. [Mehmood et al. 2024] demonstram que abordagens de aprendizado profundo para detecção de *Smishing* superam algoritmos baseados em árvores de decisão em precisão e redução de falsos positivos. Esse é um resultado interessante, dado que alertas incorretos comprometem a confiança do usuário na ferramenta. [Jain et al. 2024] propõem uma solução com regras heurísticas estáticas e baixo custo computacional. Embora essas arquiteturas sejam leves, elas são suscetíveis a variações de vocabulário usadas pelos golpistas para burlar regras estáticas, e falham em capturar a intenção semântica do discurso persuasivo independentemente das palavras exatas utilizadas.

Grandes Modelos de Linguagem e detecção semântica. A segunda geração adota modelos transformer para superar as limitações do ML clássico. [Souza et al. 2020]

introduzem o BERTimbau, pré-treinado em bilhões de palavras do corpus brWaC e da Wikipedia em português, estabelecendo o estado da arte em tarefas de PLN em português brasileiro.[Fischer et al. 2022] e [Moreira et al. 2023] aplicam BERTimbau à detecção de fake news em português, demonstrando que o ajuste fino supera significativamente abordagens baseadas em frequência de palavras em tarefas que exigem compreensão contextual. No domínio de segurança, [Ariza et al. 2022] demonstram que ataques automatizados de Engenharia Social em redes sociais permitem escalabilidade na exploração da confiança dos usuários, evidenciando a necessidade de soluções que compreendam o discurso persuasivo e não apenas palavras-chave isoladas. Contudo, essas abordagens compartilham uma limitação: o modelo classifica o risco matematicamente, mas não explica ao usuário *por que* o conteúdo oferece risco, nem oferece orientação sobre como agir.

Letramento digital. Essa frente reconhece que a detecção não é suficiente sem suporte à decisão do usuário. [Bitaab et al. 2025] demonstram que LLMs com ajuste fino superam modelos de ML clássico na detecção de fraudes ao oferecer resultados interpretáveis sobre as táticas enganosas identificadas — superando a limitação de abordagens anteriores. [Chen et al. 2025], com AI@ntiPhish, focam majoritariamente na estrutura técnica da URL e heurísticas de infraestrutura. [Tan et al. 2024], com ScamGPT-J, aplicam LLMs para simular e detectar golpes em mensagens instantâneas, destacando que usuários mais vulneráveis são frequentemente os menos receptivos a alertas técnicos.

O presente trabalho se diferencia da literatura em três frentes combinadas: 1) ao contrário das abordagens de ML clássico e detecção binária, o VeritaPlugin combina o BERTimbau com um pipeline RAG determinístico para gerar justificativas fundamentadas em enquadramento legal específico do contexto brasileiro; 2) diferente de soluções focadas em URLs e infraestrutura técnica (como AI@ntiPhish), o VeritaPlugin atua na camada semântica do discurso, permitindo detectar golpes em texto puro sem links maliciosos; 3) a integração desses elementos em uma extensão de navegador funcional e avaliada empiricamente é, até onde sabemos, inédita em português brasileiro.

3. Proposta de Solução

A Engenharia Social abrange um espectro amplo de vetores de ataque, desde *phishing* por e-mail à manipulação presencial. As redes sociais, por sua vez, constituem um ecossistema de plataformas com dinâmicas distintas. Diante dessa amplitude, este trabalho adota o Facebook como ambiente de prova de conceito: trata-se de uma plataforma com um número expressivo de usuários no Brasil e, conforme evidenciado na literatura, com alta incidência de golpes direcionados a perfis populacionais vulneráveis [Kotzias et al. 2025].

Com este foco, apresentamos a proposta do VeritaPlugin em quatro dimensões: a metodologia empírica de coleta com personas, o modelo de ameaça e as categorias de golpe detectadas, a arquitetura do sistema e as decisões de design que a motivaram, por fim, a conformidade com a LGPD como requisito arquitetural.

3.1. Metodologia de Coleta Empírica com Personas

A delimitação do escopo de detecção e a construção do dataset de calibração do modelo foram fundamentadas em uma metodologia empírica baseada em personas fictícias, criadas para estimular o algoritmo de recomendação do Facebook e coletar ocorrências reais de fraudes direcionadas a perfis populacionais distintos.

Foram criadas três personas com atributos demográficos e comportamentais contrastantes e cada persona executou uma trajetória de navegação sistemática: engajamento ativo com postagens, aceitação de solicitações de amizade e interação com mensagens diretas, com o objetivo de expor o perfil aos mecanismos de recomendação e aos agentes maliciosos presentes na plataforma.

- **Persona 1** - Antônio Ribeiro, 68 anos, aposentado. Representou usuários com letramento digital limitado e tendência a confiar em perfis que simulam autoridade institucional. Sua trajetória resultou na coleta de *advanced-fee scams*, esquemas de *phishing* bancário, *catfishing* e sorteios fraudulentos.
- **Persona 2** - Ana Lúcia, 65 anos, viúva. Modelou usuários com alta vulnerabilidade afetiva e uso intenso do Messenger. O perfil concentrou golpes de romance *baiting*, falsos consultores sentimentais, lojas falsas e *hashjacking* de *hashtags*.
- **Persona 3** - Luana Ferreira, 15 anos, estudante. Representou jovens com elevado letramento operacional mas senso crítico de segurança ainda em formação. A coleta incluiu *fake news*, manipulação por *bots, honeytraps e gaslighting*.

Durante a execução, enfrentou-se um desafio metodológico: a Persona 2 original foi banida pela plataforma e substituída por perfil equivalente, fato que corrobora a hipótese de que certos perfis demográficos são ativamente visados por agentes maliciosos.

O conjunto de ocorrências coletadas por esse processo constituiu o dataset empírico utilizado para a construção dos exemplos de *Few-Shot Learning* [Wang et al. 2020] que calibram o modelo de inferência, conferindo à solução granularidade cultural e linguística específica ao cibercrime brasileiro.

3.2. Modelo de Ameaça e Escopo de Detecção

O modelo de ameaça do VeritaPlugin parte de uma premissa central: a superfície de ataque relevante não está na camada de rede, mas na camada semântica e no discurso do agente malicioso. Alinhado à definição [NIST 2021], o sistema considera como ameaça qualquer conteúdo textual que empregue princípios de persuasão psicológica para induzir o usuário a realizar ações prejudiciais ou revelar informações confidenciais, independentemente da presença de links externos ou artefatos técnicos maliciosos. O escopo de detecção foi delimitado em seis categorias, selecionadas pela recorrência e pelo impacto social no contexto brasileiro:

- **Golpes de Ganho Financeiro Ilusório:** exploração da necessidade econômica via promessas de retorno irreal, tipificados sob o Art. 171 do Código Penal e a Lei nº 14.155/2021.
- **Golpes de Desinformação Digital:** uso de automação e bots para criar a "Ilusão de Maioria" e validar narrativas falsas por falso consenso social.
- **Fraudes em Lojas Virtuais Falsas:** preços irreais e *hashjacking* de *hashtags* populares para ampliar o alcance de anúncios fraudulentos no Marketplace.
- **Ataques de Phishing e Roubo de Dados:** indução ao fornecimento de credenciais, com risco de invasão de dispositivo (Art. 154-A do Código Penal) e tratamento ilícito de dados pessoais (Lei nº 14.155/2021).
- **Golpes Baseados em Relacionamento Romântico:** estelionato sentimental e extorsão (Art. 158 do Código Penal), com detecção da transição entre sedução e coerção financeira.

- **Mensagens Seguras:** categoria tecnicamente essencial para mitigar a Fadiga de Alertas, garantindo que conteúdos legítimos sejam reconhecidos como tal e preservando a usabilidade do sistema.

O sistema opera exclusivamente sobre texto visível no DOM, excluindo deliberadamente imagens, vídeos, áudios, malwares de sistema operacional e atribuição forense de autoria. Essa delimitação visa otimizar os recursos computacionais para o problema central e respeitar o princípio da minimização de dados da LGPD.

3.3. Arquitetura

A arquitetura do VeritaPlugin é organizada em três módulos funcionais: um módulo de captura, responsável pela extração do conteúdo do DOM sob demanda; um módulo de inferência, que realiza a classificação semântica via modelo LLM no *backend*; e um módulo de apresentação, que renderiza o resultado de forma diretamente na interface da página. A Figura 2 apresenta uma visão geral do pipeline do VeritaPlugin. Os três módulos funcionais são descritos abaixo e são representadas seis camadas do pipeline, que serão detalhadas na seção 4. A arquitetura remota foi escolhida para permitir orquestração das chamadas à OpenAI Responses API, atualização dos exemplos de *Few-Shot* sem redistribuição da extensão, e isolamento da chave de API no ambiente do usuário.

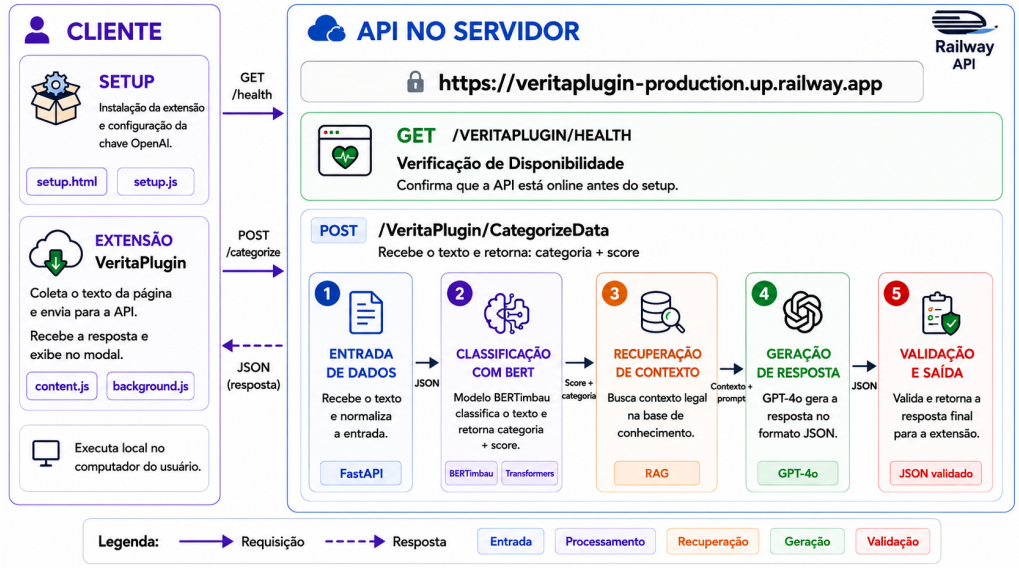

Figura 2. Visão geral do pipeline do VeritaPlugin.

Módulo de Captura. Esse módulo foi implementado como um *Content Script* seguindo as diretrizes do Manifest V3 (MV3), padrão obrigatório para extensões em navegadores baseados em Chromium (Google Chrome, Microsoft Edge). A adoção do MV3 impõe restrições arquiteturais relevantes, como a eliminação das *background pages* em favor dos *Service Workers*, determinantes para o design da solução.

O componente injeta um botão de ação (“Analisar Post”) na interface do Facebook, posicionado estrategicamente junto ao botão nativo de criação de publicação. Ao ser acionado, o sistema entra em modo de seleção: os blocos de texto extraíveis do DOM são destacados visualmente, indicando ao usuário quais elementos estão disponíveis para

análise. O usuário seleciona o trecho desejado, e o *Content Script* captura a cadeia de caracteres selecionada juntamente com a URL da página.

Módulo de Inferência. Esse módulo é responsável pela classificação semântica do texto capturado. O pipeline opera da seguinte forma: o *backend* recebe o texto e a URL via requisição HTTPS, executa a inferência junto à OpenAI Responses API e retorna ao *Content Script* uma resposta estruturada em JSON.

A escolha por *Few-Shot Learning* como estratégia de calibração foi motivada pela necessidade de sensibilidade cultural ao cibercrime brasileiro. Modelos de linguagem de propósito geral, treinados predominantemente em corpora anglófonos, apresentam limitações na identificação de nuances linguísticas regionais. Os exemplos de calibração são extraídos diretamente do dataset empírico coletado pelas personas, conferindo ao modelo referências concretas de golpes reais aplicados no contexto brasileiro e viabilizando a localização cultural da solução sem a necessidade de *fine-tuning* do modelo base.

Em caso de falha na API ou indisponibilidade do *backend*, o sistema retorna mensagens de erro amigáveis sem comprometer a aba do navegador. A chave de API da OpenAI é fornecida pelo próprio usuário durante o *onboarding* e armazenada localmente via chrome.storage.local, nunca sendo transmitida para servidores externos além da própria OpenAI. Essa decisão transfere o controle do custo de inferência ao usuário e elimina a necessidade de uma infraestrutura de gerenciamento de credenciais no *backend*.

Módulo de Apresentação. A interface do VeritaPlugin foi projetada segundo uma abordagem educativa e não autoritária, diferenciando-se dos antivírus tradicionais que bloqueiam o acesso sem explicar o risco. Ao invés de impedir a interação, o sistema sobrepõe um modal informativo que expõe os gatilhos psicológicos detectados, o enquadramento legal da ameaça e as ações defensivas recomendadas.

As decisões de design foram orientadas por heurísticas de usabilidade [Nielsen 1994]. O destaque visual dos blocos selecionáveis e a opção explícita de cancelamento a qualquer momento contribuem para controle e liberdade do usuário. O indicador de progresso durante o processamento da API mitiga a incerteza na latência de rede, contribuindo para a visibilidade do Status do Sistema. A linguagem do modal foi calibrada priorizando vocabulário acessível sem perda da informação jurídica.

3.4. Conformidade Legal

A conformidade com a Lei Geral de Proteção de Dados (Lei nº 13.709/2018) não foi tratada como uma camada de auditoria posterior ao projeto, mas como um requisito primário que moldou diretamente as escolhas de design da solução.

A decisão de restringir a análise à seleção manual faz com que o sistema não monitore o comportamento de navegação nem armazene histórico de análises, operando exclusivamente sobre o conteúdo que o usuário submete explicitamente. Isso se relaciona com o princípio da minimização de dados: o sistema processa exclusivamente o texto visível selecionado pelo usuário, sem coletar metadados de navegação, histórico de análises ou qualquer informação além do estritamente necessário para a inferência solicitada. O princípio da finalidade delimita o tratamento a um único objetivo legítimo, explicitado nos Termos de Uso apresentados no *onboarding*. O princípio do livre acesso e consentimento é operacionalizado pelo botão de acionamento explícito: a análise só

ocorre mediante manifestação inequívoca do usuário, não de forma automática.

Após o processamento da inferência, os textos analisados são descartados imediatamente, sem retenção em logs ou bancos de dados. Toda a comunicação entre o *Content Script* e o *backend* é criptografada via HTTPS/TLS, mitigando riscos de interceptação.

4. Implementação do VeritaPlugin

Os três módulos funcionais descritos na Seção 3.3 são realizados por um pipeline híbrido composto por seis camadas, desde a captura do texto pelo usuário até a entrega do resultado estruturado na interface. Informações adicionais, incluindo documentação e demonstração, estão disponíveis no site do projeto: https://verita-plugin-web-site.vercel.app/. A Figura 2 apresenta a visão geral do pipeline do VeritaPlugin. Cada camada do pipeline utiliza as seguintes tecnologias:

- Camada 1 (captura de texto) é implementada como extensão Chrome seguindo o Manifest V3, com os arquivos content.js e background.js e armazenamento local via chrome.storage.local;
- Camada 2 (entrada de dados) expõe uma API REST em FastAPI, servida pelo Uvicorn e hospedada no Railway;
- Camada 3 (classificação) utiliza o BERTimbau (neuralmind/bert-base-portuguese-cased) com ajuste fino via Hugging Face Transformers, com inferência em PyTorch CPU-only;
- Camada 4 (recuperação de contexto legal) implementa um RAG determinístico em Python puro, com recuperação por chave exata no dicionário BASE_CONHECIMENTO;
- Camada 5 (geração de resposta) utiliza o GPT-4o via API da OpenAI com Structured Outputs (strict: True);
- Camada 6 (validação e saída) realiza validação das referências legais.

4.1. Dataset BrScamsFacebook

A construção de um dataset próprio foi necessária tanto para o treinamento do modelo de classificação (Camada 3) quanto para a composição dos exemplos de calibração do pipeline de geração (Camada 5). Para isso, foram integrados oito datasets — um interno e sete externos —, totalizando 450 instâncias com textos e categorizações validadas.

Dataset Interno. O dataset interno foi construído a partir de conteúdos reais publicados no Facebook — incluindo posts, comentários e anúncios —, coletados por meio das personas descritas na Seção 3.1, cujas contas foram utilizadas para navegar e observar a plataforma em condições autênticas de uso. A definição das categorias partiu das subcategorias mapeadas na fundamentação teórica; subcategorias com baixa incidência foram agrupadas em categorias mais amplas, alinhadas aos padrões efetivamente observados no contexto brasileiro. Os dados foram estruturados com as colunas Categoria, Subcategoria, Texto, Link do post, Usuário, Perfil e Explicação. Após filtragem de qualidade, 64 instâncias foram retidas para compor o dataset.

Datasets Externos. Com o objetivo de ampliar a representatividade do dataset, foram incorporadas sete bases externas, organizadas conforme as categorias do projeto e apresentadas na Tabela 1. Os datasets externos passaram por tratamentos específicos conforme sua origem. Fontes em inglês foram submetidas a tradução automática com revisão

ID	Categoria	Fonte Principal	Instâncias
Externo 1	Golpes de Relacionamento	ScamWarners (Romance Scams)	65
Externo 2	Phishing e Roubo de Dados	Kaggle – SMS Spam Collection	77
Externo 3	Desinformação Digital	Kaggle – Fake News Portuguese	104
Externo 4	Golpes Financeiros (Tipo 1)	GitHub – sec-fraudimabr (WhatsApp/Telegram)	36
Externo 5	Golpes Financeiros (Tipo 2)	ScamWarners (Financial Scams)	25
Externo 6	Lojas Virtuais Falsas (Tipo 1)	Kaggle – Fake Reviews Dataset	43
Externo 7	Lojas Virtuais Falsas (Tipo 2)	ScamWarners (Online Selling)	36
Interno	Múltiplas categorias	Pesquisa com Personas no Facebook	64

Tabela 1. Composição dos datasets externos do BrScamsFacebook.

manual para adequação ao português brasileiro. Datasets originalmente estruturados em formatos binários foram convertidos para o formato de texto plano utilizado no pipeline. Nos casos em que a distribuição original era desbalanceada entre classes, foi realizada amostragem intencional para priorizar exemplos representativos das categorias.

A seleção das amostras foi inicialmente aleatória, seguida de curadoria manual para remoção de exemplos com baixa qualidade textual, ambiguidade semântica ou inadequação ao domínio. A distribuição final resultou em 75 instâncias por categoria — totalizando 450 exemplos com aproximadamente 16,7% por classe —, garantindo equilíbrio entre as categorias e evitando viés do modelo em favor de classes mais frequentes. Nos casos em que foi necessária tradução automática, os textos foram revisados manualmente para garantir coerência semântica e adequação ao português brasileiro; reconhece-se, contudo, que esse processo pode introduzir ruído semântico, configurando uma limitação inerente ao dataset.

4.2. Treinamento do Modelo de Classificação

Escolha do Modelo Base. Para a tarefa de classificação em português a partir do BrScamsFacebook, três abordagens foram avaliadas: TF-IDF com SVM (estabelecido como baseline, preterido por não capturar semântica contextual), GPT-3.5 com fine-tuning (descartado pelo custo computacional e menor reprodutibilidade) e BERTimbau (neuralmind/bert-base-portuguese-cased). O BERTimbau foi adotado como modelo final por já ter sido pré-treinado em português [Souza et al. 2020], conferindo-lhe sensibilidade a gírias, abreviações e construções informais típicas de redes sociais, a um custo compatível com o ambiente disponível. O ajuste fino foi realizado com Hugging Face Transformers no Google Colaboratory (GPU NVIDIA Tesla T4, 16 GB de VRAM)[2].

***Fine-Tuning* e Avaliação.** O ajuste fino adicionou uma camada de classificação ao topo do modelo pré-treinado, ajustando todos os pesos a partir das 450 instâncias rotuladas. O treinamento foi configurado para no máximo quatro épocas com *early stopping* de paciência 2 — interrompendo automaticamente caso o F1-macro no conjunto de validação não melhore por duas épocas consecutivas —, e o melhor *checkpoint* é salvo como modelo final. Os hiperparâmetros seguiram práticas consolidadas para fine-tuning de BERT com datasets reduzidos: *Learning rate: 2e-5, Batch size: 16, Épocas (máx.): 4, Weight decay: 0,01, Optimizer: AdamW, Max length: 512 tokens, Semente aleatória: 42.*

A avaliação principal utiliza validação cruzada estratificada com k=5 folds, que produz estimativas de desempenho no formato F1-macro = média ± desvio padrão —

[2]*https://colab.research.google.com/*

mais estáveis do que o holdout fixo (80/10/10), dado que com apenas 7 a 8 exemplos por categoria no conjunto de teste uma única predição errada pode alterar o F1 de uma categoria em até 14 pontos percentuais. A preparação dos splits aplica estratificação dupla por categoria e por dataset de origem, evitando que o modelo aprenda a distinguir o estilo de escrita em vez dos padrões reais dos golpes. Cabe registrar que, no k-fold, o fold de teste é utilizado simultaneamente como conjunto de validação para o *early stopping*. Com apenas 450 instâncias e 75 por categoria, reservar um terceiro split independente reduziria o conjunto de treino.Trata-se de uma limitação inerente ao tamanho do corpus, mas consistente com a prática adotada para corpora reduzidos [Souza et al. 2020].

Ao final de cada inferência, o BERTimbau retorna um score de confiança — obtido via softmax sobre os logits — categorizado em três faixas operacionais: ALTA (>0,85), MODERADA (0,60–0,84) e BAIXA (<0,60), definidas empiricamente a partir da distribuição de probabilidades no conjunto de validação.

4.3. Pipeline de Geração e API REST

A categoria predita pelo BERTimbau alimenta um pipeline RAG determinístico que injeta o contexto jurídico correspondente no prompt enviado ao GPT-4o, o qual gera a resposta estruturada em JSON com Structured Outputs (strict: True). A recuperação é feita por chave exata no dicionário BASE_CONHECIMENTO, ajudando na rastreabilidade total das referências legais para o domínio fechado de seis categorias. Uma camada de validação verifica se as leis retornadas pelo modelo pertencem ao conjunto permitido para a categoria, descartando referências inválidas antes de entregar o JSON ao *Content Script*.

Aspecto	Versão 1 (sem RAG)	Versão 2 (com RAG)	Versão 3 (RAG Aprimorado)
Modelo	GPT-5	GPT-4o	GPT-4o
Contexto injetado	Nome da categoria e lista de leis	Definição, modus operandi, análise jurídica e ações recomendadas	Idem V2
Restrição de leis	Parâmetro enum no schema JSON	Instrução no prompt + validação por substring simples	Instrução no prompt + validação por tokenização robusta (_tokens_da_lei)
Base jurídica	Leis hardcoded por categoria	BASE_CONHECIMENTO estruturado (6 categorias)	Idem V2 + LGPD e Arts. 138-140 CP para Desinformação
Categoria Seguro	Sem tratamento específico	Instrução explícita para campos vazios	Instrução explícita para campos vazios
Princípios jurídicos no prompt	Não	Parcialmente (regras de formato e não inventar leis)	Sim — PREMISSA FUNDAMENTAL, falso negativo > falso positivo, limiares por lei
Qualidade do Motivo	Genérico	Específico, com referência ao texto	Específico, consistente e juridicamente aprimorado
Structured Outputs	Instável (GPT-5)	Suporte completo com strict: True	Suporte completo com strict: True

Figura 3. Comparativo entre as três versões do prompt.

O prompt passou por três versões iterativas, sintetizadas na Figura 3. A Versão 1 utilizou contexto estático e o modelo GPT-5; incompatibilidades com Structured Outputs motivaram a migração para o GPT-4o. A Versão 2 incorporou o RAG e obteve explicações significativamente mais específicas, mas apresentou falhas na validação de leis por substring e cobertura jurídica insuficiente para desinformação. A Versão 3 corrigiu essas limitações com validação por tokenização robusta, ampliação da base jurídica de desinformação (LGPD e Arts. 138–140 CP), princípios jurídicos explícitos no prompt, incluindo a priorização de falso negativo sobre falso positivo e sanitização de texto.

5. Avaliação

5.1. Demonstração do VeritaPlugin

O VeritaPlugin é distribuído via página web e a instalação é realizada em cinco etapas: download, aceite dos Termos de Uso, verificação de disponibilidade do *backend*, inserção da chave OpenAI e confirmação de prontidão. A Figura 4 apresenta parte desse processo.

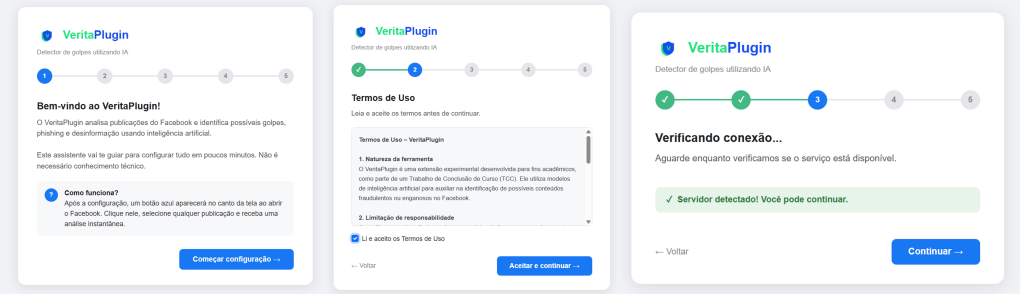

Figura 4. Processo de instalação do plugin.

Em uso, o botão "Analisar Post" é exibido fixo na interface do Facebook. Ao acioná-lo, os blocos de texto extraíveis do DOM são destacados e o usuário seleciona o conteúdo suspeito. Um *overlay* de carregamento fornece feedback visual durante o processamento; o resultado é exibido em um modal com os campos Categoria, Explicação, Base Legal e Ações Recomendadas (Figura 5).

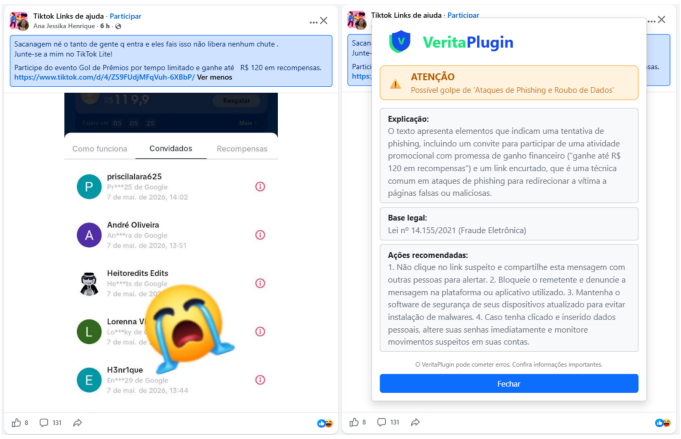

Figura 5. Exemplo do VeritaPlugin com um resultado.

5.2. Avaliação do Modelo de Classificação

A métrica principal é o F1-macro, calculada como média simples (não ponderada) do F1-score de cada categoria, na qual todas as classes contribuem igualmente independentemente do seu número de exemplos. Essa métrica penaliza igualmente o mau desempenho em qualquer classe, comportamento desejável num sistema em que falhar em qualquer tipo de golpe tem custo prático equivalente. O BERTimbau fine-tuned obteve F1-macro de 0,763 ± 0,034 na validação cruzada estratificada k=5, superando o baseline TF-IDF+SVM (0,67) em 9 pontos percentuais e excedendo amplamente o classificador por maioria. No holdout fixo, o modelo atingiu F1-macro de 0,828 e acurácia de 82,2%. A Tabela 2 apresenta o relatório consolidado.

Golpes de Relacionamento Românticо (F1 = 0,87) — Maior F1 entre as categorias. Com precisão de 0,85 e recall de 0,89, a categoria de Golpes de Relacionamento foi a mais bem classificada no k-fold consolidado. O alto *recall* indica que o modelo raramente deixa de identificar golpes românticos. O padrão linguístico dessa categoria (declarações de afeto excessivo, identidades falsas e pedidos de dinheiro após vínculo emocional) é suficientemente distinto para que o BERTimbau aprenda a reconhecê-lo com consistência.

Golpes Financeiros, Phishing e Desinformação (F1 0,78) — Desempenho Intermediário e Homogêneo. As três categorias intermediárias atingiram F1 muito próximo entre si (0,78), refletindo um padrão de confusão mútua semanticamente coerente. Golpes financeiros e phishing compartilham o uso de linguagem de urgência e solicitações de dados pessoais ou financeiros. Phishing e desinformação compartilham o uso de links maliciosos e alertas falsos. Esses padrões de confusão correspondem à sobreposição semântica real entre as categorias — o que indica que o modelo está aprendendo características linguísticas genuínas dos golpes, e não artefatos superficiais dos datasets.

Fraudes em Lojas Virtuais Falsas (F1 = 0,71) — Desempenho Abaixo da Média do Conjunto. A categoria de Lojas Virtuais Falsas apresentou a menor precisão entre as categorias de golpe (0,75) e o menor recall (0,68), resultando em F1 de 0,71. A principal fonte de confusão é a sobreposição com a categoria de Golpes Financeiros: anúncios de produtos com preços irreais frequentemente utilizam a mesma linguagem persuasiva de golpes de investimento, tornando a distinção difícil inclusive para anotadores humanos.

Categoria Seguro (F1 = 0,64) — Menor F1 entre as categorias. A categoria Seguro apresentou o menor F1 consolidado (0,64), com precisão de apenas 0,61 — indicando que 39% dos exemplos classificados como Seguro pertenciam a outras categorias. Esse resultado é esperado e metodologicamente relevante: conteúdo seguro não possui um padrão linguístico único e característico como as categorias de golpe, sendo definido pela ausência de características fraudulentas em vez da presença de padrões específicos. Essa limitação intrínseca à natureza da categoria é uma oportunidade de melhoria.

A qualidade das justificativas geradas pelo pipeline foi reforçada por dois mecanismos complementares: instrução explícita no *system prompt* para referenciar apenas leis do contexto injetado, e validação programática pós-geração que descarta referências fora do conjunto curado para cada categoria. É importante destacar que não foram observadas confusões sistemáticas entre categorias semanticamente distantes — como classificar um Golpe de Relacionamento como Phishing, ou Desinformação como Golpe Financeiro. Esse padrão valida a qualidade das representações aprendidas pelo BERTimbau: os erros ocorrem sempre entre categorias com sobreposição real de características linguísticas, e não de forma aleatória.

O desempenho obtido deve ser contextualizado frente a três fatores estruturais: o tamanho reduzido do corpus (450 instâncias), a heterogeneidade das fontes de dados — com textos originalmente em inglês e traduzidos convivendo com mensagens nativas do Facebook em português — e a natureza informal da linguagem utilizada, marcada por abreviações, erros ortográficos intencionais e construções típicas de redes sociais.

Categoria	Precisão	Recall	F1-Score	Suporte
Golpes de Relacionamento	0,85	0,89	0,87	75
Golpes Financeiros Ilusórios	0,79	0,77	0,78	75
Phishing e Roubo de Dados	0,82	0,75	0,78	75
Fraudes em Lojas Virtuais Falsas	0,75	0,68	0,71	75
Golpes de Desinformação Digital	0,77	0,80	0,78	75
Seguro	0,61	0,68	0,64	75
Acurácia geral	—		**76,00%**	**450**
F1-macro	—		**0,763**	**450**

Tabela 2. Relatório consolidado da validação cruzada.

Nesse cenário, o F1-macro de 0,763 ± 0,034 representa um resultado robusto. O desvio padrão de 0,034 é compatível com a literatura para datasets de tamanho similar, indicando que o desempenho não é artefato de uma divisão favorável dos dados. A convergência entre F1-macro e F1-weighted - esta última uma média do F1-score por categoria ponderada pelo número de instâncias de cada classe - (ambos 0,763 no consolidado) confirma o balanceamento efetivo do dataset entre as seis categorias.

6. Conclusão

Este trabalho apresentou o VeritaPlugin, uma extensão de navegador para detecção semântica de Engenharia Social no Facebook, voltada a usuários com diferentes níveis de letramento digital. Em resposta à questão de pesquisa, o classificador obteve F1-macro de 0,763 ± 0,034 na detecção das categorias de golpe, e o pipeline RAG+GPT-4o demonstrou capacidade de gerar justificativas fundamentadas.

Como limitações, reconhecemos o tamanho reduzido do corpus (450 instâncias), a dependência de tradução automática para parte dos datasets externos e o uso do *fold* de teste como conjunto de validação, que pode introduzir um viés otimista nas métricas reportadas. A categoria de conteúdo **Seguro**, com F1 de 0,64, representa a principal oportunidade de melhoria técnica na continuidade do trabalho.

Para trabalhos futuros, consideramos três direções: (i) a expansão do BrScamsFacebook com uma nova coleta e anotação nativa em português, eliminando a dependência de tradução automática e permitindo avaliação com divisão treino/validação/teste mais independente; (ii) a avaliação jurídica das justificativas geradas pelo pipeline RAG, conduzida por especialista em direito para verificar a pertinência das referências legais por categoria; (iii) e a avaliação de usabilidade com usuários reais representativos das personas definidas, para verificar se a solução cumpre seu objetivo central de proteger populações com menor letramento digital. A integração com a API da OpenAI introduz dependência de um serviço comercial externo. A avaliação de modelos alternativos de código aberto também será considerada para trabalhos futuros, visando maior autonomia da solução.

A ausência de ferramentas que atuem na camada semântica, identificada na literatura, motivou este trabalho e permanece como direção relevante para a área de segurança centrada no usuário. Avaliações futuras permitirão determinar como os resultados técnicos obtidos se traduzem em proteção real para os perfis de usuários vulnaráveis.

Disponibilidade de Dados e Uso de IA

O dataset BrScamsFacebook, para golpes digitais no contexto brasileiro, está disponível publicamente: *https://www.kaggle.com/datasets/brunaassisgt/brscamsfacebook*. Os scripts de coleta, tratamento e treinamento estão disponíveis no repositório do projeto: *https://github.com/brunanoroes/VeritaPlugin*. Ferramentas de Inteligência Artificial Generativa foram utilizadas como apoio na redação de partes do texto. Todo o conteúdo foi revisado, validado e é de responsabilidade das autoras.

Referências

Ariza, M., de Azambuja, A. J. G., Nobre, J. C., and Granville, L. Z. (2022). Ataques automatizados de engenharia social com o uso de bots em redes sociais profissionais. In *Simpósio Brasileiro de Segurança da Informação e de Sistemas Computacionais (SBSeg)*, pages 153–166. SBC.

Bitaab, M., Karimi, A., Lyu, Z., Mosallanezhad, A., Oest, A., Wang, R., Bao, T., Shoshitaishvili, Y., and Doupé, A. (2025). Scamnet: Toward explainable large language model-based fraudulent shopping website detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 27841–27848.

Bleiman, R., Park, H., and Rege, A. (2025). Educating students on the behavioral and psychological aspects of romance scam victimization via a social engineering competition. *Journal of Cybersecurity Education, Research and Practice*, 2025(1).

BRASIL (2018). Lei nº 13.709, de 14 de agosto de 2018. lei geral de proteção de dados. `https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm`.

Chen, Y.-H., Chen, J.-L., and Chiu, S.-P. (2025). Ai-enhanced phishing detection: Leveraging advanced transformer-based representations with gan for robust security. In *2025 27th International Conference on Advanced Communications Technology (ICACT)*, pages 156–161. IEEE.

Cialdini, R. (2007). Influence: the psychology of persuasion (revision edition).

Daim, T., Yalcin, H., Mermoud, A., and Mulder, V. (2024). Exploring cybertechnology standards through bibliometrics: Case of national institute of standards and technology. *World Patent Information*, 77:102278.

Ferreira, A., Coventry, L., and Lenzini, G. (2015). Principles of persuasion in social engineering and their use in phishing. In *International Conference on Human Aspects of Information Security, Privacy, and Trust*, pages 36–47. Springer.

Fischer, M., Haque, R., Stynes, P., and Pathak, P. (2022). Identifying fake news in brazilian portuguese. In *International Conference on Applications of Natural Language to Information Systems*, pages 111–118. Springer.

Frasão, A., Heinrich, T., and Fulber-Garcia, V. (2025). O cenário atual de golpes em redes sociais: uma revisão da literatura. *Computer on the beach*, 16:1–8.

Jain, A. K., Panday, A., and Goel, D. (2024). S-defender: A smishing detection approach in mobile environment. In *International Conference on Cyber Warfare, Security and Space Computing*, pages 68–78. Springer.

Jamil, A., Asif, K., Ghulam, Z., Nazir, M. K., Alam, S. M., and Ashraf, R. (2018). Mpmpa: A mitigation and prevention model for social engineering based phishing attacks on facebook. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 5040–5048. IEEE.

Kemp, S. (2025). Digital 2025: Brazil — DataReportal – Global Digital Insights. `https://datareportal.com/reports/digital-2025-brazil`.

Kotzias, P., Pachilakis, M., Aldana-Iuit, J., Caballero, J., Sánchez-Rola, I., and Bilge, L. (2025). Ctrl+ alt+ deceive: Quantifying user exposure to online scams. In *NDSS*.

Laor, T. (2022). My social network: Group differences in frequency of use, active use, and interactive use on facebook, instagram and twitter. *Technology in Society*, 68:101922.

Mehmood, M. K., Arshad, H., Alawida, M., and Mehmood, A. (2024). Enhancing smishing detection: A deep learning approach for improved accuracy and reduced false positives. *IEEE Access*, 12:137176–137193.

Mitnick, K. D. and Simon, W. L. (2003). *The art of deception: Controlling the human element of security*. John Wiley & Sons.

Moreira, L. S., Lunardi, G. M., de Oliveira Ribeiro, M., Silva, W., and Basso, F. P. (2023). A study of algorithm-based detection of fake news in brazilian election: Is bert the best. *IEEE Latin America Transactions*, 21(8):897–903.

Nascimento, A., Gadelha, T., Monteiro, J. M., and Machado, J. (2023). Uma abordagem para detecçao automática de fraudes em aplicativos de mensagens instantâneas. In *Simpósio Brasileiro de Segurança da Informação e de Sistemas Computacionais (SBSeg)*, pages 251–264. SBC.

Nielsen, J. (1994). Enhancing the explanatory power of usability heuristics. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 152–158.

NIST (2021). National institute of standards and technology - glossary. `https://csrc.nist.gov/glossary/term/Cyber\_threat`.

Silva, M. L., Teodoro, E. F., and do Couto, D. P. (2024). Psicanálise e redes sociais: as dinâmicas de captura virtual do desejo e digitalização da subjetividade nas cenas do semiocapitalismo. *Pretextos-Revista da Graduação em Psicologia da PUC Minas*, 9(17):422–442.

Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: pretrained bert models for brazilian portuguese. In *Brazilian conference on intelligent systems*, pages 403–417. Springer.

Tan, X. W., See, K., and Kok, S. (2024). Scamgpt-j: Inside the scammer's mind, a generative ai-based approach toward combating messaging scams. *arXiv preprint arXiv:2412.13528*.

Wang, Y., Yao, Q., Kwok, J. T., and Ni, L. M. (2020). Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, 53(3):1–34.