

DETECÇÃO EM CAMADAS DE ENGENHARIA SOCIAL EM AMBIENTES CORPORATIVOS: ANÁLISE LÉXICA, ESTILOMETRIA E CONFIRMAÇÃO AUTOMATIZADA

Modalidade: Resumo Expandido

Isabelli Heringer Dela Coleta*;

Leticia Siguinolfi de Lima;

* isabelliheringer@gmail.com

1 INTRODUÇÃO

Fraudes de engenharia social em ambientes corporativos, sobretudo ataques do tipo Business Email Compromise (BEC), representam um risco relevante de fraude financeira para as organizações. Sinais tradicionais de phishing, como erros de português, urgência exagerada e formatação inconsistente, tornaram-se menos confiáveis quando usados isoladamente, à medida que modelos de linguagem passaram a ser empregados na geração de conteúdo de phishing coerente e persuasivo (SIVANESWARAN et al., 2026).

Isso levanta a pergunta que orienta este trabalho: como detectar tentativas de engenharia social que combinam conteúdo persuasivo com formulações cada vez mais naturais? A hipótese é que uma abordagem em camadas, combinando análise de conteúdo, verificação de estilo e confirmação, sinaliza fraudes que escapariam de uma análise unidimensional. O objetivo é desenvolver e avaliar preliminarmente um protótipo de defesa em camadas para pequenas e médias empresas.

2 REFERENCIAL TEÓRICO

Uma revisão sistemática de 36 estudos publicados entre 2023 e 2025 mostra que modelos de linguagem facilitam a geração de phishing persuasivo e também sua detecção, e que a maioria dos estudos usa conjuntos de dados elaborados manualmente, e não bases públicas (SIVANESWARAN et al., 2026). Em português, a base PHISHING-PT (SCARAVELLI et al., 2019) reúne dados para filtragem automática de phishing em e-mails.

A Camada 2 dialoga com a literatura de estilometria e atribuição de autoria: Abbasi e Chen (2008) apresentam trabalho seminal com o sistema Writeprints, e Brocardo et al. (2013) aplicaram estilometria à verificação de autoria em mensagens curtas, abordagem simplificada aqui para o contexto corporativo em português. Na detecção de BEC, Adjei e Ayivor (2025) compararam um modelo semântico (transformer) com uma abordagem psicolinguística baseada em atributos linguísticos e estruturais e relataram maior acurácia do primeiro (F1 de 0,998 contra 0,938) e menor custo computacional do segundo. Longtchi et al. (2022) situam esses ataques no campo da engenharia social digital.

As categorias da Camada 1 combinam princípios de persuasão consagrados na literatura (CIALDINI, 2001), especialmente autoridade e escassez (esta relacionada à urgência), diretrizes do CERT.br (2026a, 2026b) e características operacionais de ataques

de BEC, como pedido de dado sensível e acesso remoto, que não correspondem diretamente a princípios de persuasão.

3 METODOLOGIA

Trata-se de pesquisa aplicada e exploratória, com avaliação quantitativa preliminar de duas das três camadas do protótipo, implementado em TypeScript/Node.js, com backend em Express, banco PostgreSQL e interface web.

A Camada 1 (detecção léxica) procura termos de cinco categorias de gatilhos psicológicos: urgência, autoridade, sigilo, pedido sensível e solicitação de acesso remoto. Cada categoria tem um vocabulário clássico (sinais óbvios) e um sofisticado (formulações mais naturais, que também podem ser produzidas com apoio de IA generativa), com cerca de 5 a 7 e de 4 a 6 expressões, respectivamente, elaborados pela autora a partir da literatura. O texto é normalizado (minúsculas, sem acentos) e comparado por expressões regulares com fronteira de palavra; variações morfológicas não previstas no vocabulário não são reconhecidas. Cada categoria detectada contribui uma única vez: 20 pontos se houver ao menos um termo clássico, ou 8 se houver apenas termos sofisticados. Soma-se um bônus de 15 pontos por categoria adicional detectada, com teto de 100. A mensagem é classificada como suspeita com 50 pontos ou mais, o que, na prática, exige duas categorias com termos clássicos ou ao menos três categorias.

A Camada 2 (verificação estilométrica) compara a mensagem ao perfil de escrita do suposto remetente, calculado como a média de quatro características em suas mensagens anteriores: palavras por frase; riqueza vocabular (palavras distintas sobre o total); formalidade (marcadores formais sobre o total de marcadores formais e informais encontrados, com valor neutro de 0,5 na ausência de ambos); e exclamações por 100 palavras. Para cada característica, calcula-se o desvio percentual em relação ao perfil, limitado a 100%. A divergência é a média dos quatro desvios, e a mensagem é considerada divergente com 40% ou mais.

A Camada 3 (protocolo de confirmação) exige confirmação por canal independente quando a Camada 1 classifica a mensagem como suspeita e ela contém as categorias pedido sensível ou acesso remoto. Nesta versão, as saídas das Camadas 1 e 2 são exibidas em conjunto ao usuário, sem regra formal de fusão: apenas a Camada 1 aciona a Camada 3.

Os pesos e o limiar da Camada 1 foram definidos de forma heurística durante o desenvolvimento, sem calibração formal; o limiar da Camada 2 foi escolhido após observar parte dos exemplos que depois integraram os testes. Os conjuntos de teste são sintéticos, gerados com apoio da ferramenta de IA generativa Claude (Anthropic) e revisados e rotulados pela autora. A Camada 1 foi avaliada com 19 mensagens (9 legítimas, incluindo uma com termo de urgência de uso legítimo, e 10 suspeitas, escritas para simular tentativas de fraude). A Camada 2 foi avaliada com 18 mensagens (9 compatíveis e 9 divergentes), distribuídas em três perfis sintéticos (casual, formal e telegráfico), cada um construído a partir de quatro mensagens, com seis mensagens de teste por perfil. A classificação do sistema foi comparada ao rótulo atribuído, tomando-se como positiva a classe suspeita (Camada 1) ou divergente (Camada 2), e calcularam-se

precisão, recall e F1 a partir de verdadeiros positivos (VP), falsos positivos (FP) e falsos negativos (FN). Não houve separação entre conjuntos de desenvolvimento e de teste.

4 RESULTADOS E DISCUSSÃO

Na Camada 1, obteve-se precisão de 100%, recall de 60% e F1 de 0,75 no conjunto completo (Tabela 1). Dos quatro falsos negativos, dois eram ataques de acesso remoto (pontuação 20 e 0) e dois eram mensagens com duas categorias detectadas apenas por termos sofisticados (31 pontos, abaixo do limiar de 50). Excluindo os dois ataques de acesso remoto (17 mensagens), o recall foi de 75% e o F1 de 0,857. Na Camada 2, obteve-se precisão de 80%, recall de 88,9% e F1 de 0,842. Os dois falsos positivos foram mensagens legítimas do perfil casual sem marcadores informais nem exclamação, o que gerou desvio máximo (100%) em formalidade e em exclamação; o falso negativo ocorreu no perfil telegráfico, em que apenas a extensão das frases diferia do perfil (divergência de 25%).

Tabela 1. Métricas preliminares por camada

Camada	n	VP	FP	FN	Precisão	Recall	F1
C1 (completo)	19	6	0	4	100%	60%	0,75
C1 (sem acesso remoto)	17	6	0	2	100%	75%	0,857
C2 (estilo)	18	8	2	1	80%	88,9%	0,842

VP: verdadeiros positivos; FP: falsos positivos; FN: falsos negativos. Fonte: elaborado pela autora (2026).

Um caso ilustra a possível complementaridade das camadas: uma mensagem obteve 20 pontos na Camada 1 (abaixo do limiar de 50, logo considerada legítima) e 52% de divergência na Camada 2 (acima do limiar de 40%), sendo sinalizada apenas por esta. Trata-se de um único exemplo e as camadas foram avaliadas em conjuntos distintos; os resultados são descritivos e não permitem concluir que o sistema combinado supera cada camada isolada.

Os resultados devem ser lidos como verificação preliminar de funcionamento, e não como estimativa de desempenho em ambientes reais. As amostras são pequenas (17 a 19 mensagens), embora a literatura revisada indique o uso frequente de conjuntos elaborados manualmente (SIVANESWARAN et al., 2026). Não houve separação entre desenvolvimento e teste: vocabulário, pesos e mensagens foram elaborados na mesma etapa, e as mensagens suspeitas partiram das mesmas categorias do vocabulário, o que tende a favorecer as métricas da Camada 1; as mensagens divergentes da Camada 2, por sua vez, foram escritas para contrastar com o perfil nas dimensões medidas pelo método. Há ainda possível viés circular, pois perfis e mensagens vieram de uma única ferramenta de IA generativa, da mesma classe das que o sistema pretende detectar. Por fim, a Camada 3 não foi avaliada quantitativamente.

5 CONSIDERAÇÕES FINAIS

O trabalho apresentou um protótipo de defesa em camadas contra engenharia social corporativa, combinando detecção léxica, verificação estilométrica e confirmação por canal independente. Em dados sintéticos e amostras pequenas, as camadas léxica e

estilométrica apresentaram desempenho compatível com a hipótese e, em um caso, detectaram sinais distintos, o que sugere complementaridade; os resultados não permitem afirmar desempenho em ambientes reais nem superioridade do sistema integrado. Como próximos passos, recomenda-se ampliar os conjuntos e separar desenvolvimento e teste, com mensagens de várias ferramentas geradoras e bases em português, como a PHISHING-PT; calibrar pesos e limiares; avaliar a Camada 3 e uma regra de decisão conjunta; e, com dados reais, observar consentimento informado, anonimização e a Lei Geral de Proteção de Dados (LGPD).

REFERÊNCIAS

- ABBASI, A.; CHEN, H. Writeprints: a stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Transactions on Information Systems*, v. 26, n. 2, 2008.
- ADJEI, Y. O.; AYIVOR, F. Semantic superiority vs. forensic efficiency: a comparative analysis of deep learning and psycholinguistics for business email compromise detection. *arXiv*, 2025. *arXiv:2511.20944*. Acesso em: 04 out. 2026.
- BROCARD, M. L.; TRAORE, I.; SAAD, S.; WOUNGANG, I. Authorship verification for short messages using stylometry. In: *INTERNATIONAL CONFERENCE ON COMPUTER, INFORMATION AND TELECOMMUNICATION SYSTEMS (CITS)*, 2013, Piraeus-Athens. *Anais [...]*. [S.l.]: IEEE, 2013. p. 1-6.
- CERT.BR. Golpes: não se deixe enganar. Cartilha de Segurança para Internet, 2026a. Disponível em: <https://cartilha.cert.br>. Acesso em: 04 out. 2026.
- CERT.BR. Golpes: evite fraudes. Cartilha de Segurança para Internet, 2026b. Disponível em: <https://cartilha.cert.br>. Acesso em: 04 out. 2026.
- CIALDINI, R. B. *Influence: science and practice*. 4. ed. Boston: Allyn & Bacon, 2001.
- LONGTCHI, T.; RODRIGUEZ, R. M.; AL-SHAWAF, L.; ATYABI, A.; XU, S. Internet-based social engineering attacks, defenses and psychology: a survey. *arXiv*, 2022. *arXiv:2203.08302*. Acesso em: 04 out. 2026.
- SCARAVELLI, L.; IKEIZUME, K.; NOGUEIRA, R.; LOCHTER, J. Phishing-pt: base de dados em português para filtragem automática de phishing em e-mail. In: *IADIS IBERO AMERICAN CONFERENCES WWW/INTERNET 2019 AND APPLIED COMPUTING 2019*. *Anais [...]*. [S.l.: s.n.], 2019. p. 109-115.
- SIVANESWARAN, D.; HEWAGE, C. T. E. R.; HERATH, H. M. K. K. M. B.; RATHORE, R. S.; SINGH, V. K.; JIANG, W. A systematic literature review of large language models in phishing attack generation and detection. *Array*, v. 30, p. 100775, 2026. DOI: 10.1016/j.array.2026.100775.