ESTUDO DE TECNOLOGIAS DE DETECÇÃO DE DEEPFAKES: USABILIDADE E CONTRIBUIÇÕES AO LETRAMENTO DIGITAL

Gabriel Santos Afini da Silva[1*], Robson Parmezan Bonidia[1], Rogério Santos Pozza[1]
[1]Universidade Tecnológica Federal do Paraná, Campus Cornélio Procópio
[*]e-mail: gsilva.2020@alunos.utfpr.edu.br

Área Temática: [5] – Ciência/Engenharia da Computação/Software

Palavras-chave: *Deepfakes*; Letramento Digital; *Chatbot*; Detecção Humana; Inteligência Artificial.

RESUMO

Contexto: O avanço das técnicas de Inteligência Artificial (IA) Generativa, especialmente as Redes Adversariais Generativas (GANs), possibilita a criação de mídias sintéticas altamente realistas, conhecidas como *deepfakes*. Esse fenômeno torna a distinção entre conteúdos reais e manipulados um desafio técnico e social de alta complexidade, exigindo novas abordagens para a preservação da integridade informacional[1]. Sob uma perspectiva sociotécnica, a desinformação deve ser compreendida como um problema estruturado em múltiplas camadas, abrangendo desde a infraestrutura tecnológica até as interfaces de usuário e os impactos sociais[2].
Lacuna/Problema: As ferramentas automatizadas de detecção apresentam limitações relacionadas à dependência de bases de treinamento específicas, à baixa capacidade de generalização diante da evolução dos modelos e à rápida obsolescência frente a novos modelos generativos, o que torna as soluções puramente algorítmicas insuficientes para o combate à desinformação [3]. Soma-se a isso o efeito *Dunning-Kruger*, no qual a falta de conhecimento pode dificultar o reconhecimento da própria limitação na identificação de conteúdos sintéticos, favorecendo uma percepção inadequada da capacidade de detecção[4]. Jovens, embora assíduos usuários de tecnologia, frequentemente dependem de heurísticas superficiais que podem falhar diante de manipulações sofisticadas[5]. Nesse contexto, o Letramento Crítico em IA constitui uma abordagem para desenvolver conhecimentos e práticas de análise crítica diante de conteúdos produzidos por IA[6].
Objetivo: O presente estudo tem como objetivo desenvolver um *chatbot* educacional baseado em princípios de Letramento Crítico em IA para auxiliar na identificação de imagens reais e sintéticas, estruturando seus conteúdos a partir de critérios diagnósticos de análise visual e estabelecendo um protocolo experimental para sua futura avaliação com estudantes ingressantes do ensino superior.
Método: O trabalho foi estruturado em duas frentes complementares. A primeira consistiu na elaboração de um protocolo experimental destinado à futura avaliação da capacidade de estudantes ingressantes no ensino superior em identificar imagens reais e sintéticas. O protocolo contempla três etapas: avaliação inicial por intuição, intervenção educacional baseada em sete critérios diagnósticos de análise visual e consolidação do aprendizado por meio de *feedback* explicativo imediato. Também foram previstos o levantamento de dados sociodemográficos e hábitos de consumo digital dos participantes, possibilitando análises posteriores sobre diferenças de desempenho entre os participantes. Para o protocolo, foi estruturado uma base de 36 imagens, sendo 21 reais e 15 sintéticas, distribuídas em níveis crescentes de dificuldade técnica. A segunda frente consistiu no desenvolvimento de um *chatbot* educacional para *Telegram*, concebido a partir dos mesmos critérios diagnósticos e destinado à disponibilização acessível dos conteúdos de letramento. Para o chatbot, foi estruturada uma base de 14 imagens sintéticas para exemplificar os critérios diagnósticos. O sistema foi desenvolvido em *Python*, com arquitetura modular por idioma, suporte aos idiomas português, inglês e espanhol e hospedagem em ambiente de nuvem. O *chatbot* constitui um artefato educacional derivado dos critérios de letramento, destinado a operacionalizar e tornar acessíveis os conteúdos desenvolvidos, não integrando o protocolo experimental.
Resultados: Foi consolidada uma base de estímulos visuais destinada ao protocolo experimental, composta por 36 imagens reais e sintéticas distribuídas em diferentes níveis de dificuldade. Também foi desenvolvido

um *chatbot* educacional para *Telegram*, com operação contínua em três idiomas: português, inglês e espanhol. A ferramenta operacionaliza os critérios de análise visual definidos no estudo por meio de orientação sobre sete aspectos: olhos, boca e dentes, orelhas, cabelo, iluminação, fundo e incoerências gerais. A interface foi estruturada para reduzir a carga cognitiva durante a consulta aos conteúdos, considerando princípios relacionados ao uso de heurísticas cognitivas na avaliação de credibilidade de informações[7]. Os artefatos técnicos, incluindo o banco de estímulos visuais e o código-fonte, foram disponibilizados em repositórios públicos, contribuindo para a transparência e a replicabilidade do trabalho. A avaliação empírica da efetividade da intervenção educacional e de seus efeitos sobre a percepção da própria capacidade de detecção permanece como etapa futura da pesquisa.
Conclusão: O estudo resultou no desenvolvimento de um *chatbot* educacional baseado em princípios de Letramento Crítico em IA e na estruturação de um protocolo experimental para investigação da capacidade humana de identificação de imagens sintéticas. Os critérios de análise visual foram transformados em orientações pedagógicas estruturadas e disponibilizados por meio de uma ferramenta acessível no *Telegram*. Embora a aplicação do experimento com participantes humanos ainda dependa da aprovação do Comitê de Ética em Pesquisa (CEP), a infraestrutura tecnológica e o delineamento metodológico foram estabelecidos. Dessa forma, a contribuição desta etapa da pesquisa concentra-se na transformação de critérios teóricos de Letramento Crítico em IA em um artefato educacional e em um protocolo experimental replicável[6]. Como trabalho futuro, prevê-se a aplicação do protocolo e a análise estatística dos dados, permitindo avaliar empiricamente a contribuição das estratégias de letramento para a identificação de imagens sintéticas.

AGRADECIMENTOS

Os autores agradecem à Universidade Tecnológica Federal do Paraná, pela participação no Programa de Voluntariado em Iniciação Científica e Tecnológica (Edital PROPPG 08/2026 — PIVICT).

REFERÊNCIAS

[1] TWOMEY, J.; CHING, D.; PETER AYLETT, M.; QUAYLE, M.; LINEHAN, C.; MURPHY, G., What Is So Deep About Deepfakes? A Multi-Disciplinary Thematic Analysis of Academic Narratives About Deepfake Technology, **IEEE Transactions on Technology and Society**, v.6, p. 64-79 (2025);
[2] QIWEI, L.; MCDONALD, A.; HAIMSON, O. L.; SCHOENEBECK, S.; GILBERT, E., The Sociotechnical Stack: Opportunities for Social Computing Research in Non-Consensual Intimate Media, **Proc. ACM Hum.-Comput. Interact.**, v.8 (2024);
[3] ALI, S.; DIPAOLA, D.; LEE, I.; SINDATO, V.; KIM, G.; BLUMOFE, R.; BREAZEAL, C., Children as creators, thinkers and citizens in an AI-driven future, **Computers and Education: Artificial Intelligence**, v.2, p. 100040 (2021);
[4] KRUGER, J.; DUNNING, D., Unskilled and Unaware of It: How Difficulties in Recognizing One's Own Incompetence Lead to Inflated Self-Assessments, **Journal of Personality and Social Psychology**, v.77, p. 1121-34 (2000);
[5] KOPS, M.; SCHITTENHELM, C.; WACHS, S., Young people and false information: A scoping review of responses, influential factors, consequences, and prevention programs, **Computers in Human Behavior**, v.169, p. 108650 (2025);
[6] VELDHUIS, A.; LO, P. Y.; KENNY, S.; ANTLE, A. N., Critical Artificial Intelligence literacy: A scoping review and framework synthesis, **International Journal of Child-Computer Interaction**, v.43, p. 100708 (2025);
[7] METZGER, M. J.; FLANAGIN, A. J., Credibility and trust of information in online environments: The use of cognitive heuristics, **Journal of Pragmatics**, v.59, p. 210-220 (2013);