Aplicação de Ferramenta Open-source na Validação de Integridade de Conteúdos Digitais para o Combate a Deepfakes Maliciosos

Thiago Oliveira Bispo de Jesus[1], **Juliana de Santi**[1], **Daniel Fernando Pigatto**[1]

[1]Universidade Tecnológica Federal do Paraná (UTFPR),
Programa de Pós-Graduação em Computação Aplicada (PPGCA)

thiagojesus@alunos.utfpr.edu.br, {jsanti,pigatto}@utfpr.edu.br

Abstract. *This paper addresses the growing problem of malicious deepfakes and proposes the use of open-source tools, such as C2PA, to verify the authenticity of digital content. The paper describes a proposed REST API to simplify the addition and validation of integrity information. The goal is to provide a transparent solution for validating and signing content, combating disinformation, with the potential for integration into decentralized social networks such as the AT Protocol, where a proof of concept will be made to reduce the spread of false information.*

Resumo. *Este artigo aborda o crescente problema dos deepfakes maliciosos e propõe o uso de ferramentas open source, como a C2PA, para verificar a autenticidade de conteúdos digitais. O trabalho descreve uma proposta de API REST para simplificar a adição e validação de informações de integridade. O objetivo é fornecer uma solução transparente para validar e assinar conteúdos, combatendo a desinformação, com potencial de integração em redes sociais descentralizadas como o AT Protocol, onde será feita uma prova de conceito para reduzir a disseminação de informações falsas.*

1. Introdução

Uma pesquisa publicada na revista digital Harvard Business Review revelou que 23% das pessoas estão utilizando inteligências artificiais (IAs) para criação de conteúdos, incluindo geração de imagens, propagandas e textos a serem postados em redes sociais [Zao-Sanders 2024].

Esse alto número de criação de conteúdos por IAs levanta questões importantes sobre segurança e privacidade. Atores maliciosos exploram essas tecnologias para realizar crimes cibernéticos, como ataques de *phishing* sofisticados e campanhas de desinformação em larga escala [Heiding et al. 2023]. Um dos exemplos mais preocupantes é o uso de *deepfakes* maliciosos, que são vídeos, imagens ou áudios manipulados para parecerem autênticos. Deepfakes podem ser usados para desinformação, manipulação de opiniões, fraudes, extorsão e ataques de engenharia social.

A constante melhoria dos algoritmos de *Large Language Model (LLM)* na criação de conteúdos torna a detecção de *deepfakes* maliciosos um desafio crescente, com necessidade de aprimoramento nos algoritmos de detecção [Farooq et al. 2025]. A acessibilidade dessas tecnologias torna a criação deste tipo de conteúdo escalável, com aplicativos

de troca de rosto ou criação de vídeos disponíveis para *smartphones* e ferramentas de código aberto, o que aumenta o volume de *deepfakes* em circulação.

Diante desse cenário, a validação da autenticidade do conteúdo digital para combater a desinformação surge como uma alternativa viável para garantir que um conteúdo não foi adulterado por terceiros. Iniciativas como a Coalizão para Proveniência e Autenticidade de Conteúdo (C2PA) [Consortium 2024] visam criar aplicações *open-source* que permitam a verificação da autenticidade de conteúdos digitais. Um exemplo de uso é quando um usuário tira uma fotografia com sua câmera habilitada para C2PA. Nesse caso, a câmera criaria um manifesto contendo algumas asserções, incluindo informações sobre a própria câmera, uma miniatura da imagem e alguns *hashes* criptográficos que vinculam a fotografia ao manifesto. Essas asserções seriam listadas na Reivindicação, que seria assinada digitalmente e, em seguida, todo o Manifesto C2PA seria incorporado ao JPEG de saída. Esse cenário pode ser visualizado na Figura 1. O Manifesto C2PA permaneceria válido indefinidamente e cada manifesto possui um identificador único.

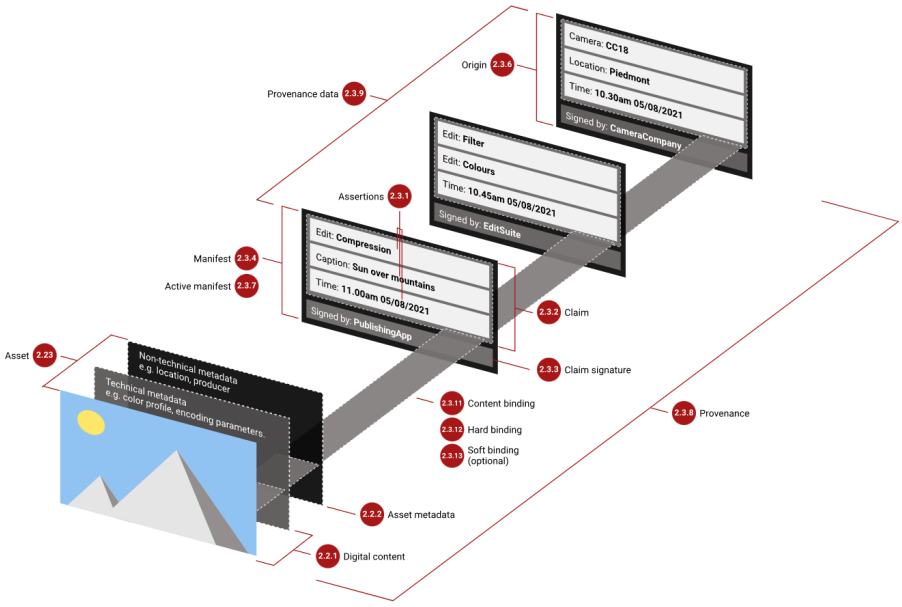

Figura 1. Cenário de incorporação de Manifesto C2PA em uma foto capturada por uma câmera [Consortium 2024].

Além disso, as redes sociais descentralizadas emergem como uma forma de mitigar os problemas de controle e manipulação presentes nas redes centralizadas. A combinação de redes sociais descentralizadas com protocolos de verificação de autenticidade tem o potencial de diminuir a distribuição de *deepfakes* e informações falsas, capacitando os usuários a terem mais controle sobre o que consomem. A colaboração entre diferentes servidores e a utilização de protocolos abertos como o AT Protocol [Kleppmann et al. 2024] possibilitam a interoperabilidade entre plataformas, criando um ecossistema social mais conectado e resistente à disseminação de informações falsas. Ao capacitar os usuários com ferramentas para verificar a autenticidade do conteúdo e descentralizar o controle sobre as informações, é possível construir um ambiente *online* mais confiável e seguro.

1.1. Objetivos

No intuito de contribuir com a iniciativa de validação de autenticidade de conteúdos digitais, é proposta a implantação de uma interface de programação de aplicação (API) [Yelavich 1985] REST [Fielding 2000] para abstrair o processo de adicionar informações de integridade, bem como a validação, tornando o processo transparente para o usuário final.

A API tem como objetivos criar e validar manifestos C2PA, suportar múltiplos formatos de conteúdo, integrar-se com plataformas existentes e fornecer maior controle ao usuário. Além de contribuir para um ecossistema de conteúdo digital mais confiável e transparente, capacitando os usuários a tomarem decisões informadas sobre o conteúdo que consomem e compartilham, a API pode ser usada como base para o desenvolvimento de novas ferramentas e aplicações que combatam a desinformação e promovam a integridade da informação *online*.

Por fim, a implantação da API será validada através de um caso de uso, com a criação de uma prova de conceito no AT Protocol [Kleppmann et al. 2024], em um fluxo de postagem de imagem com as credenciais de conteúdo embutidas na fotografia digital. Dessa forma, a implantação dessa integração entre API e uma rede social descentralizada permite a validação do processo como um todo.

2. Trabalhos relacionados

Uma publicação do *Journal of the American Society for Information Science* [Lynch 1994] cita a preocupação com a autenticidade dos conteúdos criados digitalmente a fim de manter um ambiente confiável nas redes desde os primórdios da Internet. Inúmeros trabalhos recentes abordam maneiras de garantir a autenticidade de conteúdos e combater a desinformação.

A literatura recente tem explorado múltiplas estratégias para enfrentar a desinformação e validar a proveniência de mídia em ambientes digitais. Blockchain e contratos inteligentes foram propostos por [Rashid et al. 2021] como um framework de confiança para proteção de vídeos, armazenando resumos criptográficos e metadados em redes distribuídas. Embora confiável, essa abordagem demanda infraestrutura descentralizada e não elimina a complexidade para o usuário final.

Técnicas de marcação de dados e *fingerprints* digitais [Sablayrolles et al. 2020] [Yu et al. 2021] oferecem detecção de conteúdo adversarial, introduzindo sinais imperceptíveis em *datasets* ou *outputs* de modelos generativos, possibilitando rastrear *deepfakes*, mas exigindo controle rigoroso do pipeline de treinamento.

Ferramentas de engenharia reversa e inteligência de fonte aberta (*OSINT*), como o *InVID Verification Project* [Mezaris 2018], possibilitam auditoria de imagens e vídeos por meio de busca reversa, extração de metadados e análise de *keyframes*; contudo, dependem da intervenção manual de especialistas.

Finalmente, [Hwang et al. 2021] abordam o problema pelo viés educacional, mostrando que a alfabetização midiática aumenta a resiliência social contra *deepfakes*, ainda que não ofereça mecanismos técnicos de verificação.

Esses trabalhos, embora relevantes, não oferecem uma solução integrada, automatizada e compatível com o ecossistema C2PA, lacuna que a API proposta busca preencher,

ao abstrair a criação e validação de manifestos em um fluxo transparente para usuários e plataformas.

3. Abordagem proposta

Este trabalho propõe uma API para criação e validação de manifestos C2PA, garantindo autenticidade, integridade e rastreabilidade de conteúdos digitais. A solução adota um modelo híbrido que combina assinaturas criptográficas e verificação descentralizada, permitindo a identificação indireta de *deepfakes*.

O mérito da proposta está na padronização da implementação e na redução do ônus para desenvolvedores, que podem integrar a API sem utilizar diretamente o SDK ou gerenciar chaves de assinatura. Com essa abordagem, organizações que desejam validar credenciais de conteúdo podem integrar-se à API de forma transparente, sem a necessidade de desenvolver lógica adicional, assumindo apenas a condição de confiar na solução *open-source*. A eficácia será validada em um cenário em que imagens compartilhadas em redes sociais descentralizadas incluem credenciais C2PA. Com isso, usuários poderão acessar uma publicação e navegar até um ícone de informações com o manifesto C2PA para verificar se um conteúdo é autêntico ou foi alterado antes de ser compartilhado e, dessa forma, conteúdos criados por inteligência artificial poderão ser identificados e mídias compartilhadas com intuito de desinformação não teriam o selo de autenticidade, permitindo que usuários das redes sociais as identifiquem e aí decidam se vão compartilhar.

A arquitetura do sistema, cujo esboço pode ser visualizado na Figura 2, é composta por quatro funcionalidades principais:

1. Criação de manifesto
2. Atualização de manifesto
3. Validação de assinatura pelo identificador
4. Validação de assinatura por arquivo

3.1. Comparação entre os trabalhos

Diversas abordagens existem para endereçar o mesmo problema: desinformação e compartilhamento de *deepfakes* maliciosos. Muitas vezes complementares, as abordagens diferem na forma de chegar ao resultado, demonstrando vantagens e desvantagens do seu uso dependendo do contexto. No intuito de entender para qual contexto cada trabalho apresentado poderia ser utilizado, é válida uma comparação com as características de cada solução. A Tabela 1 mostra a diferença entre as soluções propostas para combater a desinformação, em que é atribuído um número para cada um dos trabalhos relacionados:

1. [Rashid et al. 2021]
2. [Yu et al. 2021]
3. [Mezaris 2018]
4. [Hwang et al. 2021]
5. O presente trabalho

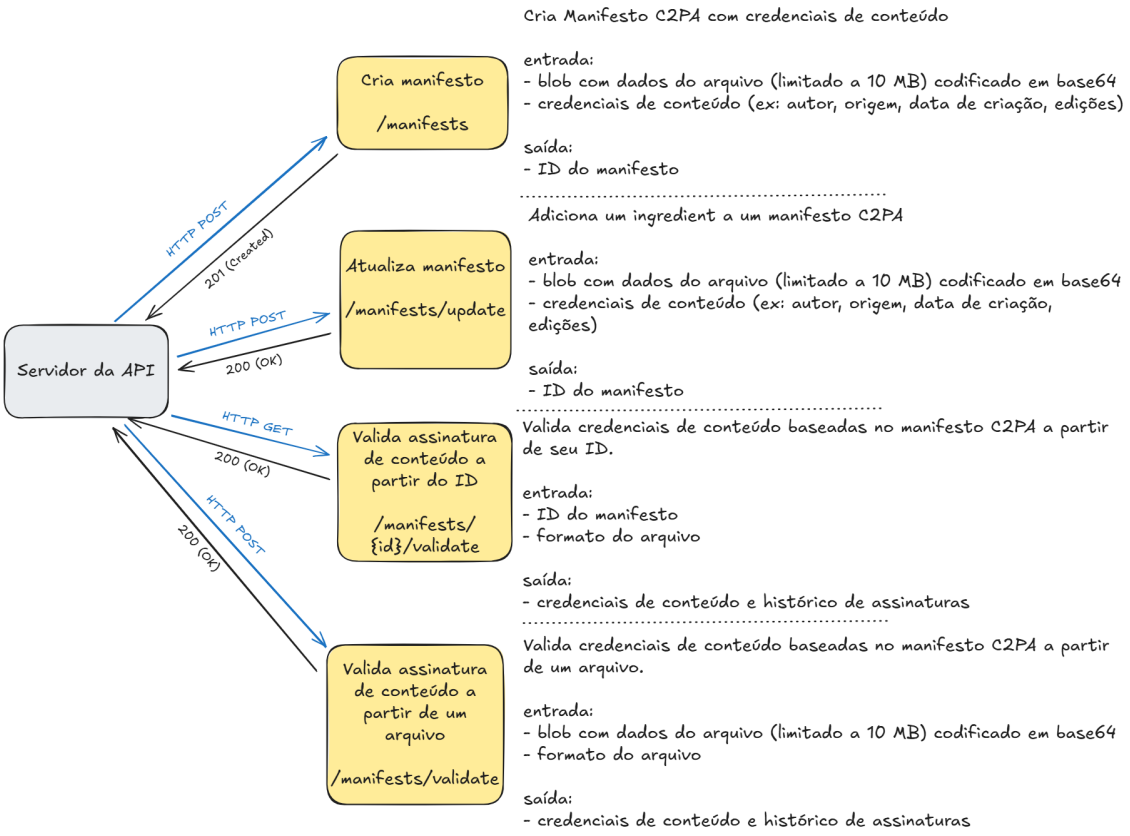

Figura 2. Esboço de API para gerenciamento das credenciais de conteúdo [Autoria própria]

Característica	1	2	3	4	5
Detecção de Deepfakes		✓	✓		✓
Validação de integridade de conteúdo		✓			✓
Rastreamento de alterações em conteúdo	✓		✓		✓
Atuação em conteúdos em diversos formatos, além de imagens e vídeos		✓		✓	✓
Solução técnica para combater desinformação	✓	✓	✓	✓	✓
Independência de serviços externos ou da capacidade cognitiva do usuário	✓	✓			✓
Abordagem atemporal, independente dos avanços tecnológicos de *deepfakes*				✓	✓

Tabela 1. Comparação entre trabalhos para combater desinformação por *deepfakes*.

4. Resultados preliminares e Trabalhos Futuros

4.1. Implementação da API

A API proposta implementa quatro funcionalidades principais: criação de manifesto, atualização de manifesto, validação por identificador e validação por arquivo. Os testes iniciais foram realizados localmente com chaves de exemplo e imagens armazenadas em uma pasta de *uploads*. As requisições foram feitas utilizando uma ferramenta de teste de APIs, validando a capacidade da API de gerar e embutir credenciais C2PA em imagens

e, em seguida, verificar sua autenticidade. O código-fonte da implementação, incluindo instruções para reprodução dos experimentos, está disponível em [de Jesus 2025a].

Os experimentos confirmam que a API é capaz de embutir credenciais C2PA em imagens e validá-las de forma transparente. A Figura 3 mostra o fluxo de criação e validação de manifestos em um único exemplo de imagem.

Figura 3. Fluxo de criação e validação de manifestos C2PA (a) Imagem original, (b) Manifesto criado, (c) Validação do manifesto pelo arquivo. [Autoria própria]

4.2. Demonstração no AT Protocol

A integração da API ao cliente da rede social AT Protocol também foi um sucesso e permitiu a exibição das credenciais de conteúdo C2PA diretamente na interface da rede social,

em que o código-fonte pode ser visto em [de Jesus 2025b]. A Figura 4 ilustra o funcionamento: (a) uma publicação com ícone de credenciais disponível, (b) a validação bem-sucedida exibindo detalhes do manifesto embutido, e (c) um exemplo de imagem sem credenciais. Essa demonstração confirma a viabilidade de incorporar verificação de autenticidade em redes descentralizadas, facilitando a identificação de conteúdos confiáveis.

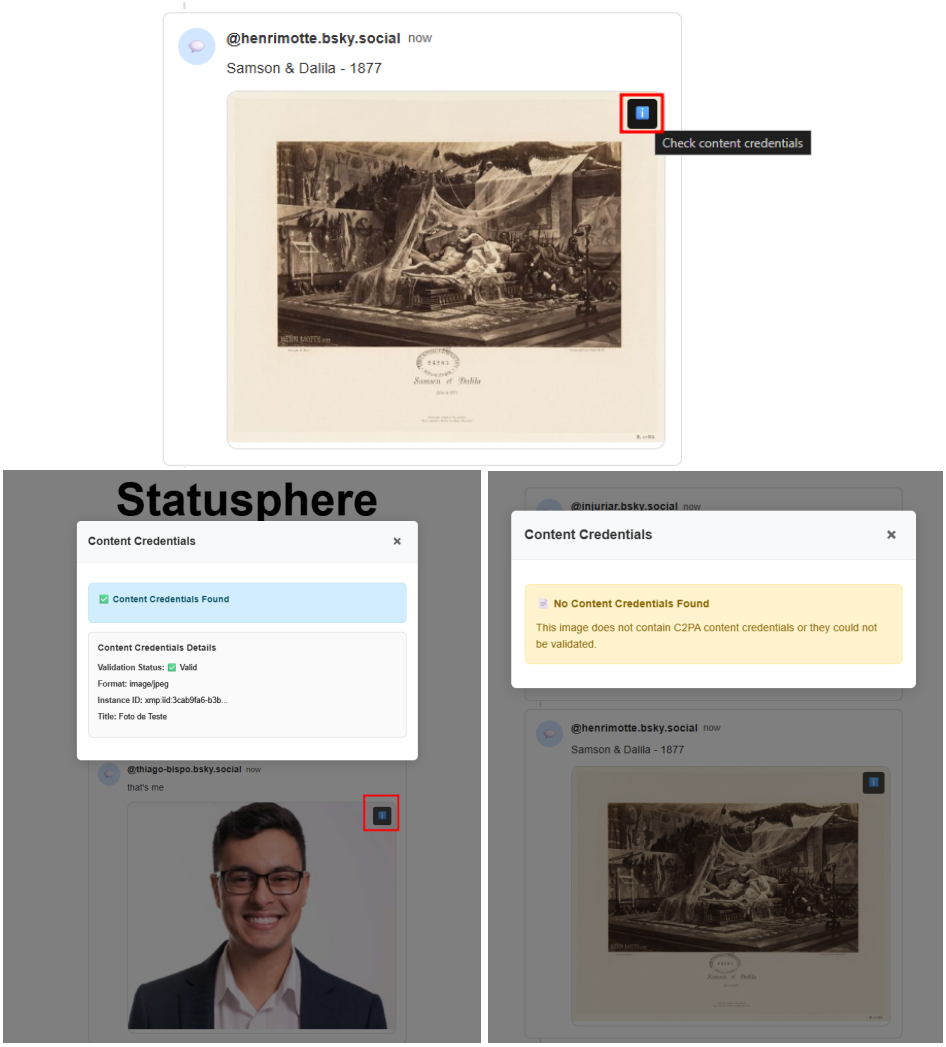

Figura 4. Demonstração de credenciais de conteúdo no Statusphere: (a) Publicação com ícone de verificação; (b) Credenciais C2PA válidas; (c) Imagem sem credenciais. [Autoria própria]

4.3. Próximos passos

Como próximos passos, planeja-se a implantação da API em uma infraestrutura escalável na nuvem, utilizando a Amazon Web Services (AWS) [Amazon Web Services 2025]. O objetivo é disponibilizar os serviços de criação e validação de manifestos C2PA de forma acessível e com alta disponibilidade, permitindo que diferentes aplicações clientes integrem-se à API sem necessidade de configuração local. Essa abordagem possibilitará balanceamento de carga, elasticidade e monitoramento contínuo, garantindo que a solução seja capaz de atender múltiplos usuários simultaneamente e sirva como base para futuras integrações com redes sociais descentralizadas em produção.

Referências

[Amazon Web Services 2025] Amazon Web Services (2025). Amazon Web Services (AWS) Documentation. Acesso em: 3 ago. 2025.

[Consortium 2024] Consortium, C. C. (2024). C2pa technical specification. `https://c2pa.org`. Acesso em: 15 mar. 2025.

[de Jesus 2025a] de Jesus, T. O. B. (2025a). C2PA API. `https://github.com/khellwan/c2pa-api`. Acesso em: 30 jul. 2025.

[de Jesus 2025b] de Jesus, T. O. B. (2025b). Statusphere C2PA. `https://github.com/khellwan/statusphere-c2pa`. Acesso em: 01 ago. 2025.

[Farooq et al. 2025] Farooq, M. U., Javed, A., Malik, K. M., and Raza, M. A. (2025). A lightweight and interpretable deepfakes detection framework.

[Fielding 2000] Fielding, R. T. (2000). *Architectural Styles and the Design of Network-based Software Architectures*. PhD thesis, University of California, Irvine. Acesso em: 26 jan. 2025.

[Heiding et al. 2023] Heiding, F., Schneier, B., Vishwanath, A., Bernstein, J., and Park, P. S. (2023). Devising and detecting phishing: Large language models vs. smaller human models.

[Hwang et al. 2021] Hwang, Y., Ryu, J. Y., and Jeong, S.-H. (2021). Effects of disinformation using deepfake: The protective effect of media literacy education. *Cyberpsychology, Behavior, and Social Networking*, 24(3):188–193.

[Kleppmann et al. 2024] Kleppmann, M., Frazee, P., Gold, J., Graber, J., Holmgren, D., Ivy, D., Johnson, J., Newbold, B., and Volpert, J. (2024). Bluesky and the at protocol: Usable decentralized social media. In *Proceedings of the ACM Conext-2024 Workshop on the Decentralization of the Internet (DIN '24)*, pages 1–9. ACM.

[Lynch 1994] Lynch, C. A. (1994). The integrity of digital information: Mechanics and definitional issues. *Journal of the American Society for Information Science*, 45(10):737–744.

[Mezaris 2018] Mezaris, V. (2018). Invid verification project. `https://www.invid-project.eu/`. Online: acesso em 15-Março-2025.

[Rashid et al. 2021] Rashid, M. M., Lee, S.-H., and Kwon, K.-R. (2021). Blockchain technology for combating deepfake and protect video/image integrity. *Journal of Korea Multimedia Society*, 24:1044–1058.

[Sablayrolles et al. 2020] Sablayrolles, A., Douze, M., Schmid, C., and Jégou, H. (2020). Radioactive data: tracing through training.

[Yelavich 1985] Yelavich, B. M. (1985). Customer information control system—evolving system facility. *IBM Systems Journal*, 24(3.4):264–278.

[Yu et al. 2021] Yu, N., Skripniuk, V., Abdelnabi, S., and Fritz, M. (2021). Artificial fingerprinting for generative models: Rooting deepfake attribution in training data. In *ICCV*.

[Zao-Sanders 2024] Zao-Sanders, M. (2024). How people are really using genai. *Harvard Business Review*.