

Desafios técnicos e riscos técnicos e jurídicos do controle digital

Autor(a): Raquel Cavalcanti Ramos Machado Malenchini

Publicado em: 12 de março de 2024

Publicação: migalhas_import

Link JurisTube: <https://juristube.com.br/colunistas/raquel-cavalcanti-ramos-machado-malenchini/desafios-tecnicos-e-riscos-tecnicos-e-juridicos-do-controle-digital>

Fonte original:

<https://www.migalhas.com.br/depeso/403214/desafios-tecnicos-e-riscos-tecnicos-e-juridicos-do-controle-digital>

Identificador citável: juristube.com.br/colunistas/raquel-cavalcanti-ramos-machado-malenchini/desafios-tecnicos-e-riscos-tecnicos-e-juridicos-do-controle-digital#ep-9c9277e1

Resumo

Por Raquel Cavalcanti Malenchini, Rodrigo Gurgel Fernandes Távora. Os impactos da expressão online levantam questões sobre limites e controle. A autorregulação regulada, onde plataformas seguem ...

Texto integral

Os impactos de postagens no ambiente digital sobre a psique humana têm levado a contínuos questionamentos sobre a liberdade de expressão e seu possível controle (quando é possível controlar, até que ponto, e quem pode fazê-lo). As reflexões envolvem compreender os limites de atuação dos usuários, das plataformas digitais, de empresas e de governos. O conteúdo da internet pode ser moderado pela regulação por terceiros (Estado), pela autorregulação ou pela autorregulação regulada. Nesta última, a plataforma modera o conteúdo com base em diretrizes definidas pelo Estado, com possível conciliação entre o interesse público e a expertise das plataformas. Por essa aparente vantagem, iniciativas legislativas têm buscado essa abordagem. Entretanto, para evitar riscos, como o uso político, o efeito silenciador ou ataques sistêmicos, o sucesso da autorregulação regulada depende de uma discussão profunda e de uma especificação detalhada. A moderação pode ocorrer por iniciativa da própria plataforma, ou seja, uma autorregulação interna, ou decorrer de uma notificação externa. A implementação da autorregulação regulada impõe desafios não apenas legislativos, mas técnicos, como o custo da infraestrutura e o acerto das decisões tomadas por IA, ainda que com o apoio humano. Na autorregulação interna, por exemplo, o uso de máquinas não é suficiente para realizar essa tarefa, e uma revisão por equipes treinadas é necessária. Mesmo na detecção de material de abuso infantil, como o Child Sexual Abuse Material - CSAM, em que o reconhecimento e padrões de imagens permite a identificação de conteúdos inadequados, a intervenção humana é essencial. Na Google, por exemplo, a identificação de material infantil abusivo ocorre em duas etapas: primeiro, com tecnologias automáticas, pela aplicação de buscas de correspondência de HASH 1 e por IA, seguida da revisão, por amostragem, por equipes treinadas, dada a impossibilidade de revisar todo o conteúdo detectado (JASPER, 2022). Mesmo passando por essas duas etapas, erros na identificação ocorrem. Fotos normais tiradas por pais de crianças e postadas na internet já foram erroneamente classificadas como material abusivo (MULLIN, 2022). Em um estudo da Facebook, de 150 contas denunciadas por

CSAM, descobriu-se que 75% haviam enviado imagens não maliciosas. Na análise do LinkedIn, de 75 contas denunciadas pelo software de CSAM PhotoDNA, apenas em 31 dos casos confirmou-se o CSAM. Segundo a Comissão da UE, Ylva Johnasson, no atual estágio tecnológico, a acurácia dos algoritmos de CSAM é de 88%, antes da revisão humana. O receio dos efeitos de erros deste tipo (falso positivo) é acentuado em regimes autoritários ou democracias em retrocesso. Para a regulação de conteúdo político, a dificuldade de treinar algoritmos ou de padronizar a análise de revisão é agravada pela subjetividade da classificação de conteúdo dessa natureza, muitas vezes ambíguo. Um estudo de Lim (LIM, 2018) avaliou a confiabilidade dos vereditos das agências de checagem de fatos. Observou-se uma concordância entre essas agências FactChecker e Polifact em 49 notícias, de um total de 77, com classificações que variam de " verdade " a " nitidamente falso ". Embora tenham alcançado bom desempenho em falsidades óbvias, a concordância foi baixa em afirmações ambíguas, destacando os desafios, mesmo para jornalistas, da verificação de fatos políticos. A ambiguidade na classificação de conteúdo afeta a avaliação do desempenho de sistemas automáticos de autorregulação. A contabilização precisa de acertos e erros na classificação de conteúdo político é inviável, dado que revisar todo o conteúdo é impossível. Além disso, inexiste uma referência de classificação definitiva e qualquer revisão é também subjetiva. Logo, há o risco de aplicação de sanções desproporcionais contra provedores, caso não se reconheça a falibilidade de sistemas de classificação ou exija desempenho elevado, cuja avaliação também é complexa. O receio de sanções pode levar a um efeito de silenciamento, resultando na remoção indevida de conteúdo legítimo ao ajustar os parâmetros de classificação do sistema automático de detecção. Isso é evidenciado pela análise da Curva Característica de Operação do Receptor 2 . Para diminuir os Falsos Negativos, ou seja, o conteúdo falso não identificado, a sensibilidade dos algoritmos seria aumentada, o que inevitavelmente levaria a um aumento dos Falsos Positivos, onde conteúdo verdadeiro é erroneamente considerado falso. A complexidade computacional e o desempenho dos métodos de detecção automática de fake-news são aspectos críticos. Um estudo (Hamed et al., 2023) descreve os dois modelos teóricos de detecção de fake-news: 1) Checagem de fatos baseada no conhecimento, que envolve pesquisa em um universo aberto de informações sem treinamento prévio do software; e 2) Checagem baseada em atributos, utilizando métodos supervisionados de IA treinados com bases de conteúdo rotulado. Esses atributos podem incluir informações da mensagem (como título e corpo da notícia) ou características do contexto social, como atributos de rede (padrões de difusão em redes sociais) e atributos de usuário (como credibilidade, seguidores, dados de conta e comportamento). Os modelos baseados em atributos de conteúdo, em geral, mostram os melhores resultados. O desenvolvimento desses sistemas utiliza, para treinamento e teste, conjuntos de conteúdos rotulados de dados provenientes de agências de checagem. Há uma forte correlação positiva entre o tamanho dessas bases de dados e o desempenho do sistema, destacando a importância de bases com muitas amostras. Conforme apontado por Hamed (HAMED et al. 2023), a escassez de dados é um grande obstáculo na detecção de notícias falsas, especialmente quando as bases são pequenas ou desatualizadas, não representando adequadamente a variedade das notícias. A diversidade dos rótulos na classificação também dificulta a formação de bases maiores de treinamento e teste pela fusão de diferentes bases, pois algumas usam rótulos binários e outras classificações mais refinadas. Uma alternativa seria utilizar conjuntos de dados multilíngues, como proposto em (DEMENTIEVA et al. 2023) para conteúdos globais. No entanto, para notícias políticas locais, e.g. acerca de uma eleição municipal, é essencial treinar os modelos com dados das agências de checagem locais. Isso ressalta a importância dessas agências no desenvolvimento de sistemas automáticos de detecção de fake-news. Além disso, é crucial escolher bases de dados

confiáveis para evitar vieses nos sistemas de verificação, selecionando aquelas comprovadamente apartidárias, financeiramente independentes e desvinculadas do governo, com equipes profissionais comprometidas com a ética e acreditadas por organismos de transparência. O artigo 19 do Regulamento de Serviços Digitais do Parlamento Europeu - DSA exemplifica isso ao impor requisitos de transparência, competência e experiência para as Trusted Flaggers, entidades notificadoras de conteúdo abusivo. Em 10/10/23, uma análise foi conduzida nas bases das principais agências de checagem de fatos nacionais: LUPA, Aos Fatos, Comprova e Agence France-Presse. A maior base de dados, da LUPA, continha apenas 5.742 notícias rotuladas, abrangendo vários temas como política, eleições, saúde, ciência e meio ambiente. Em comparação, a base americana Polifact, amplamente utilizada em treinamento de modelos de pesquisa, especialmente para notícias políticas, contava com mais de 20 mil amostras. A unificação das bases nacionais é desafiadora devido aos diferentes rótulos utilizados. Bases tão pequenas são inadequadas para o treinamento de sistemas automáticos de classificação em cenários de aplicações reais. Hamed et al. (2023) revisaram 17 métodos de classificação automática de fake-news textuais, desenvolvidos entre 2019 e 2023. Dentre esses sistemas, selecionamos os de Elhadad et al. (2019), Kumar (2020) e Choudhury (2023), que utilizaram, pelo menos em parte, bases de checagem de notícias políticas, como LIAR, FakeNewsNet e PolitiFact. A acurácia do sistema mais recente (Choudhury, 2023), baseado apenas na LIAR, foi de apenas 60%. No entanto, considerando possíveis vieses na análise de um único sistema, como o uso de bases pequenas, calculamos a média das acurácias desses sistemas, obtendo 67%, ainda insatisfatória e pouco acima do desempenho de um classificador aleatório (50%). Em Hoes et al. (2023), o uso do ChatGPT 3.5 na identificação de fake-news textuais resultou em uma acurácia de 69%. Esses resultados refletem o desafio atual da classificação automática de conteúdo político, ainda com resultados insatisfatórios. As Fake News podem incluir áudios e vídeos, além de texto, como as Deep Fakes - DF, que são conteúdos de imagem ou áudio realistas gerados por Deep Learning. Com o surgimento das DF, o interesse em desenvolver um padrão aberto de Autenticação Ativa de mídias ressurgiu, exemplificado pela Coalizão para Procedência e Autenticação de Conteúdo - C2PA, formada por empresas como Adobe, Microsoft, Intel e BBC. No entanto, os modelos de detecção atuais ainda aplicam a Autenticação Passiva, pela análise dos conteúdos das mídias. O principal problema destes modelos, treinados e testados em uma mesma base, é a incapacidade de generalização. Desempenhos bons para uma base de teste específica não necessariamente seriam observados em uma aplicação real. Uma abordagem mais realista de avaliação pode ser feita através de torneios, onde competidores geram mídias falsas para construir bases de teste e, em seguida, testam seus modelos pré-treinados entre si. Em Altuncu et al. (2022), uma meta-revisão de métodos propostos em torneios para detecção de vídeos falsos é apresentada. No torneio Deepfake Detection Challenge - DFDC/20, o melhor modelo entre 2.114 participantes alcançou uma acurácia de apenas 65,18%. Em relação aos áudios, conforme a revisão feita em (Yi, 2022), no torneio Audio Deep synthesis Detection - ADD, que inclui trilhas com áudios mais realistas com inserção de ruído ou contendo apenas um trecho falso, a melhor Taxa de Erro Equivalente - EER foi de 21,7%, ou seja, taxas de Falso Positivo e Falso Negativo em elevado patamar. Esses resultados mostram as dificuldades dos sistemas atuais de detecção de DF em vídeo ou áudio, com desempenhos ainda insuficientes para aplicações reais como a autorregulação regulada. Como foi observado, a acurácia dos sistemas de classificação propostos recentemente, tanto para fake-news textuais quanto para deep-fakes, é ainda baixa. A implantação de sistemas de classificação de conteúdo nessas condições, por imposição legal com sanções severas, provavelmente induziria um efeito silenciador, com a aplicação indevida de medidas de moderação

como a remoção de conteúdo. A evolução desses sistemas a um nível satisfatório para aplicação real também é incerta. Para os sistemas de classificação de conteúdo textual, a melhoria da acurácia está limitada pelo desempenho obtido por equipes de jornalistas, que, como vimos, também divergem na classificação de conteúdo político. A outra limitação é a dependência de bases de dados, que dificilmente alcançarão uma cobertura e detalhamento de informações suficiente para a classificação de notícias locais, como uma notícia sobre uma eleição municipal. Para os sistemas de detecção de deep fakes é provável que observemos melhorias de desempenho significativas no futuro. Entretanto, considerando o surgimento de novas abordagens de geração de DF, sobretudo como o uso de Redes Adversárias Generativas - GAN, a evolução do desempenho dos sistemas de detecção pode ser lenta e inconstante. É possível que uma solução definitiva surja apenas através do desenvolvimento de padrões de Autenticação Ativa, tanto para conteúdos autênticos registrados por equipamentos gravadores, como para deep fakes criadas por sistemas de IA.

----- 1 HASH é a transformação de uma grande quantidade de dados em uma pequena quantidade de informações, que permite a identificação rápida e a verificação de integridade. 2 É uma representação gráfica do desempenho de um sistema classificador binário, cujos valores de FP (Falso Positivo) e FN (Falso Negativo) variam à medida que parâmetros ou limiares de discriminação são ajustados. ----- ALTUNCU, E.; FRANQUEIRA, V. N. L.; LI, S. 2022. Deepfake: Definitions, Performance Metrics and Standards, Datasets and Benchmarks, and a Meta-Review. Arxiv, p. 1-31. CHOUDHURY, D. et al. 2023. A novel approach to fake news detection in social networks using genetic algorithm applying machine learning classifiers. DEMENTIEVA, D.; KUIMOV, M.; PANCHENKO. 2023. A. Multiverse: Multilingual Evidence for Fake News Detection. J. Imaging, vol. 9, n.4, p. 77. ELHADAD, M.K. et al. 2019. A novel approach for selecting hybrid features from online news textual metadata for fake news detection, in: International Conference on P2P, Parallel, Grid, Cloud and Internet Computing, Springer. HAMED, S.; AZIZ, M. J.; YAAKUB, M. R. 2023. A review of fake news detection approaches: a critical analysis of relevant studies and highlighting key challenges associated with the dataset, feature representation, and data fusion. Heliyon, v. 9, n. 1, p. 1-23. HOES, Emma et al. 2023. Leveraging ChatGPT for Efficient Fact-Checking. Psyarchiv Preprints, p. 1-16. JASPER, S. How we detect, remove and report. child sexual abuse material. 2022. Google. KUMAR, S. et al. 2020. Fake news detection using deep learning models: A novel approach. LIM, C. 2018. Checking how fact-checkers check. Research & Politics, v. 5, n. 3, p. 205. MARANHÃO, J.; CAMPOS, R.; GUEDES, J.; OLIVEIRA, S. R.; GRINGS, Maria Gabriela. 2021. Regulação de "Fake News" no Brasil. Instituto Legal Grounds. MULLIN, J. 2022. Google's Scans of Private Photos Led to False Accusations of Child Abuse. Electronic Frontier Foundation. YI, J. 2022. ADD 2022: the first audio deep synthesis detection challenge. ICASSP, p. 9216-9220.

Como citar este documento

As citações abaixo seguem os padrões ABNT NBR 6023:2018, APA 7ª edição, Chicago Author-Date (17ª ed.) e BibTeX (LaTeX). Copie o formato exigido pelo periódico ou gerenciador bibliográfico (Zotero, Mendeley, EndNote).

ABNT (NBR 6023:2018)

MALENCHINI, Raquel Cavalcanti Ramos Machado. Desafios técnicos e riscos técnicos e jurídicos do controle digital. *migalhas_import*, 12 mar. 2024. Disponível em: <https://www.migalhas.com.br/depeso/403214/desafios-tecnicos-e-riscos-tecnicos-e-juridicos-do-controle-digital>. Acesso via: JurisTube — Acervo Digital de Direito. Disponível em: <https://juristube.com.br/colunistas/raquel-cavalcanti-ramos-machado-malenchini/desafios-tecnicos-e-riscos-tecnicos-e-juridicos-do-controle-digital>. Acesso em: 2 out. 2026.

APA 7ª edição

Malenchini, R. C. R. M. (2024, March 12). Desafios técnicos e riscos técnicos e jurídicos do controle digital. **migalhas_import**. <https://www.migalhas.com.br/depeso/403214/desafios-tecnicos-e-riscos-tecnicos-e-juridicos-do-controle-digital>

Chicago Author-Date (17ª ed.)

MALENCHINI, Raquel Cavalcanti Ramos Machado. 2024. “Desafios técnicos e riscos técnicos e jurídicos do controle digital.” **migalhas_import**, March 12. <https://www.migalhas.com.br/depeso/403214/desafios-tecnicos-e-riscos-tecnicos-e-juridicos-do-controle-digital>

BibTeX

```
@article{raquel-cavalcanti-ramos-machado-malenchini-desafios-t-cnicos-e-riscos-t-cnicos-e-ju-2024,
author = {Malenchini, Raquel Cavalcanti Ramos Machado},
title = {Desafios técnicos e riscos técnicos e jurídicos do controle digital},
journal = {migalhas_import},
year = {2024},
url = {https://www.migalhas.com.br/depeso/403214/desafios-tecnicos-e-riscos-tecnicos-e-juridicos-do-controle-digital},
urldate = {2024-03-12}
}
```



Licença de Uso — CC BY-NC-SA 4.0

Este conteúdo está licenciado sob a licença Creative Commons Atribuição-NãoComercial-CompartilhaIgual 4.0 Internacional (CC BY-NC-SA 4.0). O uso para fins acadêmicos e educacionais é permitido, desde que citada a fonte original (JurisTube) conforme as normas técnicas indicadas neste documento.

Termos completos:

<https://creativecommons.org/licenses/by-nc-sa/4.0/deed.pt-br>

JurisTube — Acervo Digital de Direito

<https://juristube.com.br/colunistas/raquel-cavalcanti-ramos-machado-malenchini/desafios-tecnicos-e-riscos-tecnicos-e-juridicos-do-controle-digital>

PDF gerado em 2026-10-02 02:45 UTC