Da curadoria algorítmica à inteligência coletiva: proposta de observatório de comunicação científica no YouTube[1]

Lucas Augusto Ferreira OLIVEIRA[2]
Adriana C. Omena SANTOS[3]
Universidade Federal de Uberlândia – UFU

Resumo
Diante de desafios como a desinformação e a sobrecarga informacional, este estudo sobre Comunicação Pública da Ciência postula a Curadoria de Conteúdo, humana e algorítmica, como prática fundamental para mitigá-los. Para tanto, desenvolve-se um observatório de dados abertos que analisa conteúdos audiovisuais científicos em língua portuguesa publicados no YouTube. A plataforma utiliza aprendizado de máquina para processar vídeos, comentários e dados contextuais, realizando análises semânticas, de sentimento e de inferência. O objetivo é mapear ecossistemas discursivos e oferecer uma ferramenta pública que fomente a inteligência coletiva sobre a dinâmica entre ciência e sociedade.

Palavras-chave: comunicação pública da ciência; inteligência artificial; curadoria de conteúdo; youtube; dados abertos.

Introdução

A Comunicação Pública da Ciência (CPC) transcende o modelo tradicional de mera disseminação de informações, configurando-se como um processo multifacetado orientado ao engajamento e ao empoderamento cívico. Essa prática, contudo, ocorre em um contexto de profundas transformações tecnológicas que, ao apresentarem simultaneamente desafios e oportunidades, exigem de pesquisadores e profissionais a aquisição contínua de novas competências. Tal exigência evidencia-se particularmente nos ambientes digitais hiperconectados, onde publicações científicas disputam atenção com opiniões individuais em plataformas mediadas por lógicas algorítmicas. Essa conjuntura de pluralidade enunciativa, como aponta Farnese (2023), delineia um cenário complexo no qual as apropriações midiáticas se configuram como frentes decisivas de atuação.

[1] Trabalho apresentado no GP Tecnologias e Culturas Digitais, do 25º Encontro dos Grupos de Pesquisas em Comunicação, evento componente do 48º Congresso Brasileiro de Ciências da Comunicação.

[2] Doutorando em Educação e Mestre em Tecnologias, Comunicação e Educação pela Universidade Federal de Uberlândia (FACED/UFU). Bacharel em Comunicação Social pela Universidade Estadual Paulista (FAAC/Unesp). E-mail: lucasafoliveira@gmail.com

[3] Orientadora da pesquisa. Doutora em Comunicação pela ECA/USP, com pós-doutorado pela University of Ottawa. Bolsista produtividade CNPq e professora do Curso de Jornalismo, do Programa de Pós-graduação em Tecnologias, Comunicação e Educação (PPGCE) e do Programa de Pós-graduação em Educação (PPGED), todos na UFU. E-mail adriomena@gmail.com

Ademais, o avanço da Inteligência Artificial (IA) em múltiplas aplicações tem impactado significativamente as interações sociais e, por conseguinte, intensificado a intrincada dinâmica da CPC. Desafios como o excesso informacional, a proliferação de desinformação e a formação de câmaras de eco emergem como obstáculos significativos que, segundo Pereira (2023), comprometem a preservação da confiabilidade e da acessibilidade do conhecimento científico junto à sociedade. Essa problemática não pode ser desvinculada das lógicas subjacentes que governam os ambientes digitais. Sob o modo de produção capitalista, a tecnologia atua como um princípio organizador que visa elevar a produtividade e acelerar a rotação do capital, fazendo com que as plataformas operem, simultaneamente, como tecnologias de produção, comunicação e circulação (Grohmann, 2020, p. 4). Diante disso, torna-se nevrálgica uma abordagem que integre regulação e governança de dados, tratando a IA como ferramenta a ser gerida criticamente para mitigar riscos de ordem ética e social, como a perpetuação de vieses algorítmicos e a exacerbação de desigualdades.

Em face de tais desafios, a Curadoria de Conteúdo emerge como prática interdisciplinar fundamental na gestão do ecossistema digital, cujo propósito central consiste em atenuar a sobrecarga de dados mediante a seleção, organização, contextualização e compartilhamento de material relevante, visando facilitar seu acesso, apropriação crítica e reutilização. A superação desses obstáculos, contudo, não reside em uma ação centralizada. Conforme apontam Miranda *et al.* (2024), a solução aponta para o fomento de uma inteligência coletiva, como a curadoria social, sob o pressuposto de que nenhuma entidade isolada detém a capacidade de monitorar e validar a totalidade do conteúdo disseminado em múltiplas plataformas. Essa fragmentação da autoridade informacional é corroborada por Malcher *et al.* (2025, p. 308), que destacam que as redes sociais, ao assumirem papel central na distribuição de conteúdo, promovem um "deslocamento de poderes institucionais, ainda que a confiança do público esteja fortemente ancorada no conhecimento especializado".

Tal conjuntura torna imperativo que as práticas de CPC transcendam a intuição e se fundamentem em abordagens estratégicas, apropriando-se das particularidades da nova ordem digital. Essa necessidade de ação planejada ressoa nas propostas de Malcher *et al.* (2025, p. 301 e 311), que, para mitigar os danos decorrentes do excesso informacional e da baixa credibilidade da informação, apontam para caminhos experimentais como a criação de canais diretos entre cientistas e sociedade, a produção de conteúdo validado e

adaptado às dinâmicas das plataformas e o fomento ao engajamento que incentive a participação cidadã. Dentre essas estratégias, o audiovisual emerge com especial relevância, dada a sua natureza multimodal que, ao articular múltiplos recursos semióticos, permite recontextualizar o discurso científico (Luzón, 2019, p. 170, tradução nossa), tornando-o acessível a públicos heterogêneos. Essa proeminência manifesta-se com particular intensidade em plataformas como o YouTube, ambiente onde, paradoxalmente, a complexidade da comunicação científica se aprofunda. Embora a divulgação científica seja uma prática consolidada nesse espaço, a pesquisa acadêmica sobre o fenômeno ainda se encontra "em seus estágios iniciais" (Velho; Barata, 2020, p. 1, tradução nossa), evidenciando uma lacuna analítica diante de um cenário ambivalente. Tal dualidade reside no fato de as plataformas atuarem, simultaneamente, como espaços para "uma comunicação pública da ciência engajadora" e como "focos para a disseminação de desinformação" (Velho; Mendes; Azevedo, 2020, p. 1 e 11, tradução nossa).

A inter-relação desses elementos – o paradoxo das plataformas, o papel do agente comunicador e a relevância dos formatos – sublinha, portanto, a urgência de investigações empíricas que avancem para além das análises já existentes, a fim de compreender as complexas interações entre produção e consumo de conteúdo científico nesses ecossistemas. A abordagem analítica de tais ambientes exige, portanto, que se transcenda a visão meramente instrumentalista, compreendendo a relação entre capitalismo e tecnologia não como um nexo causal determinista, mas como uma complexa articulação de forças. Nessa perspectiva, as plataformas são concebidas como infraestruturas de conexão alimentadas por dados e algoritmos, nas quais estão "valores e normas inscritos em suas arquiteturas e interfaces", o que evidencia os distintos mecanismos de extração de valor que lhes são inerentes (Grohmann, 2020, p. 2).

É precisamente em resposta a essa criticidade que o presente estudo se debruça sobre as contradições e potencialidades da CPC na era da IA. A hipótese central sustenta que a Curadoria de Conteúdo, compreendida como uma mediação dialética, opera sobre a contradição fundamental entre a dimensão humana (a práxis crítica) e a dimensão algorítmica (as forças produtivas condicionadas pelo capital). Postula-se que, ao articular essa tensão, a curadoria se constitui como uma ferramenta contra-hegemônica, capaz de mitigar os efeitos da sobrecarga informacional e da desinformação. Para investigar essa hipótese, o objetivo prático desta tese consiste no desenvolvimento de um observatório

de dados abertos, focado em conteúdo audiovisual sobre ciência em língua portuguesa veiculado no YouTube. Mais do que um repositório, a plataforma proposta é concebida como uma ferramenta analítica que utiliza processamento de dados e aprendizado de máquina para gerar análises quanti-qualitativas aprofundadas. Com isso, busca-se não apenas mapear os ecossistemas discursivos e as dinâmicas de engajamento, mas também oferecer um instrumento público capaz de fomentar a inteligência coletiva e a competência crítica em informação. Em última análise, a pesquisa visa contribuir com um arcabouço metodológico e tecnológico para a análise da comunicação científica em ambientes plataformizados, oferecendo subsídios para que pesquisadores, comunicadores e a sociedade possam compreender e navegar de forma mais consciente e estratégica neste complexo cenário informacional.

Percurso metodológico

A plataforma em desenvolvimento transcende a concepção de um mero repositório de dados, configurando-se como um laboratório dinâmico, projetado para a análise crítica e aprofundada das complexas interações dialéticas entre ciência, sociedade e tecnologia. O objetivo é consolidar um conjunto de dados extenso, robusto e multidimensional, visando superar a superficialidade das métricas convencionais – como contagens de visualizações ou curtidas – para, em contrapartida, desvelar as estruturas e dinâmicas subjacentes que governam os processos de produção, circulação e recepção de conteúdo científico, com particular enfoque no audiovisual.

O desenvolvimento iniciou-se com a elaboração de um código de retaguarda modular, destinado à coleta, ao processamento e à análise de dados. Para a implementação do sistema, optou-se pela linguagem de programação Python (versão 3.12.10), reconhecida por sua sintaxe clara e seu caráter de código aberto. Como ambiente de desenvolvimento, utilizou-se o Visual Studio Code (versão 1.101.1), um editor de código-fonte versátil desenvolvido pela Microsoft. O equipamento empregado nos estágios iniciais foi um MacBook Air M1, dotado de CPU de 8 núcleos, 8GB de memória RAM e GPU integrada de 7 núcleos com suporte a Metal 3. O desenvolvimento também contou com o acesso a grandes modelos de linguagem disponibilizados pela Hugging Face, uma comunidade que facilita a colaboração em projetos de IA (Hugging Face, 2025). A escolha do canal Nerdologia no YouTube como objeto para os testes iniciais justificou-

se por sua significativa popularidade (3,4 milhões de inscritos) e pelo extenso volume de vídeos publicados (YouTube, 2025).

A primeira etapa consistiu na obtenção e otimização de modelos pré-treinados, os quais foram convertidos para formatos de alto desempenho, como o float16, com o objetivo de reduzir o consumo de memória e acelerar a inferência. Os modelos selecionados foram: Whisper MLX LM, para a transcrição automática de fala (Hannun *et al.*, 2025); Helsinki-NLP, empregado na tradução (Tiedemann *et al.*, 2022); BERTweet, especializado na análise de sentimentos em textos (Pérez; Giudici; Luque, 2021); WikiNEuRal, utilizado no reconhecimento de entidades nomeadas, tais como pessoas, organizações e locais (Tedeschi *et al.*, 2021); BGE-M3, para a criação de representações vetoriais capazes de capturar o significado semântico (Chen *et al.*, 2024); Jina Reranker, para o refinamento da relevância em buscas contextuais (Jina AI, 2024); e mDeBERTa, aplicado em tarefas de inferência de linguagem natural, permitindo avaliar relações lógicas entre frases (Laurer *et al.*, 2023).

Na segunda etapa, procedeu-se à coleta de dados brutos com o objetivo de construir um panorama informacional abrangente. Esse processo teve início com a extração sistemática de dados do canal-alvo, utilizando a API do YouTube (Google, 2025). Foram coletados metadados dos vídeos, bem como o conjunto integral de comentários e respostas associados às publicações veiculadas entre 01/07/2024 e 31/12/2024. Em seguida, realizou-se a obtenção dos áudios dos vídeos selecionados, os quais foram posteriormente processados pelo modelo de transcrição. Essa abordagem foi adotada devido à sua capacidade de gerar transcrições de alta fidelidade, superando as limitações e os erros frequentemente observados nas legendas automáticas fornecidas pelo próprio YouTube.

Posteriormente, com o objetivo de investigar as condições materiais e históricas que poderiam influenciar a elaboração de conteúdo científico audiovisual, foram agregadas informações contextuais. Dados sobre filmes e programas televisivos foram extraídos do The Movie Database (TMDB) – um banco de dados colaborativo e de código aberto – com o intuito de identificar possíveis reflexos da cultura de entretenimento (The Movie Database, 2025). Adicionalmente, consultou-se o Google para obter notícias publicadas em época similar, visando correlacionar a criação dos vídeos com acontecimentos correntes. Por fim, um acervo com mais de 40 mil ocorrências históricas, catalogado por data e natureza (como marcos, nascimentos ou óbitos de figuras públicas),

oriundo do Zenodo e originalmente em inglês, foi traduzido para o português e integrado ao sistema (Lopez; Fernandez, 2023). Essa abordagem cronológica de efemérides permite correlacionar os conteúdos audiovisuais com fatos relevantes, buscando-se, assim, compreender como a delimitação temática pode ser influenciada por fatores contextuais mais amplos.

A terceira etapa concentrou-se no processamento e na análise semântica dos dados. Inicialmente, o modelo WikiNEuRal foi aplicado a todos os insumos textuais para o reconhecimento de entidades nomeadas, identificando e classificando sistematicamente menções a fim de mapear os principais atores e conceitos discutidos. Em seguida, a análise de sentimentos, conduzida pelo modelo BERTweet, processou os comentários e respostas dos usuários, classificando-os em categorias como positivo, negativo ou neutro. Essa análise proporcionou uma avaliação quanti-qualitativa da recepção do público aos vídeos, permitindo identificar polarizações e compreender a valência afetiva dos debates gerados. Por fim, a vetorização de conteúdo, realizada pelo modelo BGE-M3, converteu as informações em representações vetoriais: densas para capturar similaridade semântica e esparsas para otimizar a busca lexical. Este procedimento é fundamental para a análise computacional, pois traduz o significado intrínseco do texto para um formato matemático, viabilizando operações complexas como buscas avançadas, agrupamento temático e medições de similaridade textual em larga escala.

A quarta e última etapa, de convergência, integrou todas as camadas de dados previamente processadas, visando gerar novas descobertas e compreensões sobre o objeto de estudo. Inicialmente, realizou-se o agrupamento de vídeos por meio de um algoritmo aglomerativo, para o qual se calculou uma matriz de similaridade combinada, ponderando as afinidades semântica, lexical e dos títulos. Em seguida, aplicou-se uma máscara temporal para assegurar que apenas vídeos publicados em datas próximas fossem agrupados, resultando na identificação de conjuntos temáticos em períodos específicos. Para cada conjunto formado, o sistema buscou itens contextuais relevantes – como filmes, programas televisivos, notícias e eventos históricos. A relevância desses itens foi então determinada por um processo multifásico: primeiramente, selecionaram-se elementos dentro de uma janela temporal compatível com o conjunto; posteriormente, calculou-se uma pontuação combinada para identificar os mais promissores; e, por fim, o modelo Jina Reranker refinou a lista, assegurando máxima pertinência.

Como procedimento final, para cada conjunto de vídeos e seu respectivo contexto, realizou-se a inferência de linguagem natural com o intuito de: (a) gerar uma premissa – uma frase que hipotetiza uma conexão causal ou temática (exemplo: "A publicação do vídeo 'A História da Penicilina' foi motivada pelo evento histórico 'Descoberta da Penicilina por Fleming', ocorrido em data próxima"); e (b) testar essa hipótese com o modelo mDeBERTa, avaliando se a premissa constituía uma inferência lógica suportada pelos dados, uma contradição, ou se a relação era neutra. Essa técnica foi adotada com o intuito de, em vez de limitar a identificação de correlações estatísticas, buscá-las ativamente, gerando informações interpretáveis que podem ser posteriormente validadas e aprofundadas.

Reflexões

Esta proposta constitui-se, em essência, como um instrumento metodológico para a análise crítica dos ecossistemas discursivos da CPC. Seu propósito central é habilitar e fomentar novas investigações sobre como temas científicos são construídos, ressignificados e propagados socialmente, servindo assim de contraponto empírico às dinâmicas muitas vezes opacas das plataformas digitais. Para materializar esses objetivos, será desenvolvida uma interface de usuário pública e interativa, projetada para democratizar o acesso aos dados e às análises geradas. Por meio desta ferramenta, um espectro diverso de atores sociais – incluindo pesquisadores, jornalistas, educadores e o público em geral – poderá explorar o acervo de forma autônoma, superando as barreiras técnicas que atualmente limitam investigações em larga escala.

A interface permitirá que os usuários não apenas naveguem e filtrem um vasto conjunto de informações – de vídeos e comentários a notícias e eventos históricos –, mas também visualizem correlações em gráficos interativos e analisem agrupamentos temáticos. De modo ainda mais fundamental, a funcionalidade de exportação de dados em formatos abertos visa catalisar análises independentes e a replicação de estudos, nutrindo um ecossistema de ciência aberta. Por meio dessa arquitetura, o observatório transcende sua função de mera ferramenta para se tornar um nó em uma rede de construção de inteligência coletiva. Em última instância, espera-se que a contribuição deste trabalho seja dupla: oferecer um diagnóstico preciso do cenário atual e, ao mesmo

tempo, legar uma infraestrutura que capacite a comunidade a monitorar, compreender e intervir de forma mais estratégica nos debates públicos sobre ciência e tecnologia.

Referências

CHEN, J. *et al.* **BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation**. arXiv Cornell University, 5 fev. 2024. Disponível em: https://doi.org/10.48550/arXiv.2402.03216. Acesso em: 21 jun. 2025.

FARNESE, P. Comunicação organizacional em universidades públicas: os desafios de comunicar a ciência na sociedade midiatizada. **Journal of Science Communication América Latina**, Trieste, v. 6, n. 01, 29 maio 2023. DOI: 10.22323/3.06010206. Disponível em: https://jcomal.sissa.it/article/pubid/JCOMAL_0601_2023_A06/. Acesso em: 7 out. 2024.

GOOGLE. **YouTube Data API**. Mountain View, 2025. Portal: Google for Developers. Disponível em: https://developers.google.com/youtube/v3. Acesso em: 20 jun. 2025.

GROHMANN, R. A Comunicação na Circulação do Capital em Contexto de Plataformização. **Liinc em Revista**, Rio de Janeiro, v. 16, n. 1, 30 maio 2020. DOI: 10.18617/liinc.v16i1.5145. Disponível em: https://revista.ibict.br/liinc/article/view/5145. Acesso em: 10 jul. 2025.

HANNUN, A. *et al.* **MLX: Efficient and flexible machine learning on Apple silicon**. Versão 0.26.1. 4 jun. 2025. Python. Disponível em: https://github.com/ml-explore/mlx.

HUGGING FACE. **Hugging Face**. [*S. l.*], 2025. Portal: The AI community building the future. Disponível em: https://huggingface.co/. Acesso em: 20 jun. 2025.

JINA AI. **Jina Reranker V2 Base Multilingual Model Card**. Sunnyvale: Jina AI, 2024. Disponível em: https://jina.ai/models/jina-reranker-v2-base-multilingual/. Acesso em: 20 jun. 2025.

LAURER, M. *et al.* Less Annotating, More Classifying: Addressing the Data Scarcity Issue of Supervised Machine Learning with Deep Transfer Learning and BERT-NLI. **Political Analysis**, Cambridge, v. 32, n. 1, p. 84–100, 9 jun. 2023. DOI: 10.1017/pan.2023.20.

LOPEZ, J. A. G.; FERNANDEZ, A. B. **WorldEvents**. Zenodo, 10 nov. 2023. Disponível em: https://zenodo.org/records/10105219. Acesso em: 22 jun. 2025.

LUZÓN, M. J. Bridging the gap between experts and publics: the role of multimodality in disseminating research in online videos. **DOAJ (DOAJ: Directory of Open Access Journals)**, [*s. l.*], n. 37, 1 maio 2019. Disponível em: https://revistaiberica.org/index.php/iberica/article/view/114. Acesso em: 9 jul. 2025.

MALCHER, M. A. *et al.* Interatividade, confiabilidade e engajamento: 22 anos de estudos sobre comunicação pública da ciência em redes sociais on-line. **Observatorio**

(OBS*), Lisboa, v. 19, n. 1, 31 mar. 2025. DOI: 10.15847/obsobs19120252616. Disponível em: https://obs.obercom.pt/index.php/obs/article/view/2616. Acesso em: 10 jul. 2025.

MIRANDA, M. K. F. de O. *et al.* A Curadoria Social e a Competência Crítica em Informação como pressupostos de combate à desinformação: um estudo de caso no YouTube. **Informação & Informação**, Londrina, v. 28, n. 2, p. 180–206, 3 maio 2024. DOI: 10.5433/1981-8920.2023v28n2p180.

PEREIRA, M. **Ciência, sociedade, divulgação científica: a visão dos cientistas**. 2023. Dissertação (Mestrado) — Universidade Federal de Minas Gerais, Faculdade de Filosofia e Ciências Humanas, Belo Horizonte, 2023.

PÉREZ, J. M.; GIUDICI, J. C.; LUQUE, F. M. Pysentimiento: A Python Toolkit for Sentiment Analysis and SocialNLP tasks. **CoRR**, [*s. l.*], abs/2106.09462, 2021. Disponível em: https://arxiv.org/abs/2106.09462.

TEDESCHI, S. *et al.* WikiNEuRal: Combined Neural and Knowledge-based Silver Data Creation for Multilingual NER. *In*: EMNLP 2021, 2021, Punta Cana. **Findings of the Association for Computational Linguistics**. Punta Cana: Association for Computational Linguistics, 1 jan. 2021. DOI: 10.18653/v1/2021.findings-emnlp.215. Disponível em: https://aclanthology.org/2021.findings-emnlp.215/.

THE MOVIE DATABASE. **TMDB API Documentation**. 2025. Disponível em: https://developer.themoviedb.org/docs. Acesso em: 20 jun. 2025.

TIEDEMANN, J. *et al.* **Democratizing neural machine translation with OPUS-MT**. arXiv Cornell University, 4 dez. 2022. Disponível em: https://doi.org/10.48550/arXiv.2212.01936. Acesso em: 20 jun. 2025.

VELHO, R. M.; BARATA, G. Profiles, Challenges, and Motivations of Science YouTubers. **Frontiers in Communication**, Texas, v. 5, 17 nov. 2020. DOI: 10.3389/fcomm.2020.542936. Disponível em: https://www.frontiersin.org/journals/communication/articles/10.3389/fcomm.2020.542936/full. Acesso em: 10 jul. 2025.

VELHO, R. M.; MENDES, A. M. F.; AZEVEDO, C. Communicating Science With YouTube Videos: How Nine Factors Relate to and Affect Video Views. **Frontiers in Communication**, Texas, v. 5, 25 set. 2020. DOI: 10.3389/fcomm.2020.567606. Disponível em: https://www.frontiersin.org/journals/communication/articles/10.3389/fcomm.2020.567606/full. Acesso em: 10 jul. 2025.

YOUTUBE. **Nerdologia**. Mountain View, 2025. Portal: YouTube. Disponível em: https://www.youtube.com/channel/UClu474HMt895mVxZdlIHXEA. Acesso em: 20 jun. 2025.