

Desvendando Comportamentos Inautênticos Coordenados no YouTube:

**Uma Abordagem de Aprendizado de Máquina para
Detectar Manipulação e Seu Impacto nos Algoritmos
de Recomendação no Brasil**

Chamada CNPq/MCTI/FNDCT N° 22/2024

Programa Conhecimento Brasil

Apoio a Projetos em Rede com Pesquisadores Brasileiros no Exterior

Rede de Pesquisadores:

Rose Marie Santini - Universidade Federal do Rio de Janeiro, Brasil

Mathias Felipe de Lima Santos - Macquarie University, Austrália

Francisco Quevedo Camargo - University of Exeter, Inglaterra

Rio de Janeiro, Brasil, 2024

Resumo

Este projeto tem como objetivo investigar o comportamento inautêntico coordenado (Coordinated Inauthentic Behavior ou CIB) nos comentários do YouTube desenvolvendo métodos de aprendizado de máquina para análises com foco na identificação de padrões de manipulação, análise do impacto do CIB nos algoritmos de recomendação da plataforma e exploração da amplificação de narrativas específicas por meio de engajamento artificial. O estudo é especialmente relevante no contexto brasileiro, onde o YouTube é uma das principais plataformas de consumo de informação audiovisual. Por meio de uma combinação de técnicas de aprendizado de máquina e processamento de linguagem natural (PLN), este projeto desenvolverá e implementará métodos para detectar CIB nas seções de comentários do YouTube. Os resultados fornecerão uma compreensão mais profunda sobre como o CIB opera, além de oferecer recomendações para mitigar sua influência na opinião pública e nos algoritmos da plataforma. Os desdobramentos incluirão novos insights sobre manipulação digital, publicações acadêmicas e recomendações práticas para plataformas e formuladores de políticas no combate à desinformação, visando manter a integridade do discurso online.

Palavras-Chave

Comportamento Inautêntico Coordenado, YouTube, Aprendizado de Máquina, Processamento de Linguagem Natural, Manipulação Algorítmica, Algoritmos de Recomendação

Abstract

This project aims to investigate Coordinated Inauthentic Behavior (CIB) in YouTube comments by developing machine learning methods and focusing on identifying patterns of manipulation, analyzing the impact of CIB on YouTube's recommendation algorithms, and exploring the amplification of specific narratives through artificial engagement. The study is particularly relevant in the Brazilian context, where YouTube is one of the main platforms for audiovisual information consumption. Through a combination of machine learning (ML) and natural language processing (NLP) techniques, this project will develop and implement methods to detect CIB in YouTube's comment sections. The findings will not only provide a deeper understanding of how CIB operates but also offer recommendations for mitigating its influence on public opinion and platform algorithms. The outcomes will include new insights into digital manipulation, academic publications, and practical recommendations for platforms and policymakers to counter disinformation and maintain the integrity of online discourse.

Keywords

Coordinated Inauthentic Behavior, YouTube, Machine Learning, Natural Language Processing, Algorithm Manipulation, Recommendation Algorithms

I – Identificação da proposta

| | |
|---|--|
| Título do Projeto | Desvendando Comportamentos Inautênticos Coordenados no YouTube: Uma Abordagem de IA para Detectar Manipulação e Seu Impacto nos Algoritmos de Recomendação no Brasil |
| Título do Projeto em Inglês | Unveiling Coordinated Inauthentic Behavior on YouTube: An AI Approach to Detecting Manipulation and Its Impact on Recommendation Algorithms in Brazil |
| Área de conhecimento predominante a áreas do conhecimento correlatas | Ciência da Informação |
| Coordenador Brasil | Rose Marie Santini |
| Instituição Executora/UF | Universidade Federal do Rio de Janeiro, Brasil |
| Coordenadores no Exterior | Mathias Felipe de Lima Santos Francisco Quevedo Camargo |
| Instituições Participantes | Macquarie University, Austrália University of Exeter, Inglaterra |
| Prazo de execução (meses) | 24 meses |
| Valor total | R\$ 600.000,00 |

II – Detalhamento do projeto

1. Introdução e discussão teórica

As plataformas de mídias sociais mudaram drasticamente o panorama da comunicação global, permitindo interações instantâneas e envolvimento em uma escala nunca vista antes. Plataformas como Facebook, Twitter/X e Instagram tornaram-se arenas onde ideias, opiniões e informações fluem de maneira desenfreada, conectando pessoas de todos os cantos do mundo. Contudo, ao mesmo tempo em que ampliaram o acesso à informação, essas

plataformas também se tornaram um campo fértil para a manipulação digital. No centro dessa nova realidade está o comportamento inautêntico coordenado (CIB), um fenômeno onde múltiplas contas, frequentemente falsas ou automatizadas, são usadas de forma deliberada para amplificar certas mensagens, influenciar o discurso público e manipular opiniões de maneira sutil, mas poderosa (de-Lima-Santos & Ceron, 2021a). Este fenômeno tem se tornado um foco crescente de pesquisa acadêmica devido ao seu impacto significativo no discurso público e nos processos políticos (Giglietto et al., 2020a; 2020b).

É nesse contexto que o YouTube, a maior plataforma de compartilhamento de vídeos do mundo, assume um papel central, especialmente em países como o Brasil, onde o consumo de conteúdo audiovisual online é uma das principais formas de acesso à informação e entretenimento (Bernardi et al., 2022, Hale et al., 2024, Kantar IBOPE Media, 2024). Com milhões de usuários ativos no país, o YouTube tornou-se uma ferramenta fundamental não só para influenciadores e criadores de conteúdo, mas também para atores mal-intencionados que se utilizam do CIB para manipular a opinião pública (Rocha et al., 2024).

No Brasil, um dos maiores mercados de YouTube do mundo, as seções de comentários tornaram-se palco de sofisticadas redes de robôs (*bots*) e contas falsas, organizadas para inflar artificialmente o engajamento de vídeos e promover narrativas específicas (Alberto et al., 2015). Esses comentários automatizados não apenas influenciam a percepção pública, mas também podem manipular os algoritmos de recomendação do YouTube, que utilizam sinais de engajamento como número de curtidas, visualizações e comentários para sugerir conteúdo a outros usuários (McCrosky & Geurkink, 2021). Ao inflar artificialmente esses números, as campanhas de CIB podem aumentar a probabilidade de que vídeos manipulados ganhem visibilidade na plataforma, afetando diretamente o que milhões de brasileiros assistem diariamente.

Seja em temas políticos, de saúde pública ou em torno de figuras públicas, essas campanhas coordenadas moldam discussões online e impactam diretamente o debate público. O CIB no YouTube, particularmente nos comentários, reflete um problema maior: a manipulação da atenção e a criação de câmaras de eco, onde os usuários são expostos a visões polarizadas e reforçam preconceitos preexistentes (Möller, 2021). Dado o impacto do YouTube no Brasil, estudar como o CIB opera nessa plataforma é crucial para compreender as dinâmicas digitais no país e seu efeito sobre o discurso público.

Além disso, o volume de informações compartilhadas em redes sociais como o YouTube impossibilita qualquer análise manual sobre CIB. A utilização de modelos de aprendizado de máquinas (*machine learning* – ML) e inteligência artificial (IA) para identificar comportamentos inautênticos em redes sociais ainda é um desafio aberto, em especial em estudos sobre o impacto de *bots* para desinformação (Schuchard et al, 2019a; 2019b). Diversos modelos desenvolvidos para essas aplicações enfrentam dificuldades em manter baixas taxas de falso positivos e evitar vieses sistemáticos ([González-Bailón & Paltoglou, 2015](#); Rauchfleisch & Kaiser, 2020; TeBlunthuis et al., 2024). O sucesso da modelagem de perfis inautênticos em redes sociais depende do correto entendimento, mitigação e contabilização (Ntoutsi et al., 2020) das diversas fontes de vieses encontradas durante o seu desenvolvimento, como fontes de dados e anotadores humanos (Hovy & Prabhumoye, 2021).

1.1 Definição e evolução do CIB

O conceito de CIB ganhou destaque em 2018, quando plataformas como Facebook começaram a enfrentar os esforços coordenados para manipular seus espaços (Giglietto et al., 2020a). O CIB geralmente envolve múltiplos atores trabalhando juntos para espalhar mensagens específicas com a intenção de influenciar a opinião pública. Esses atores podem ser contas automatizadas ou casos de coordenação utilizando humanos (Felipe, 2022). Em plataformas como o YouTube, essas campanhas coordenadas são projetadas para amplificar determinadas mensagens utilizando contas falsas, robôs e até participantes humanos que se envolvem em comportamentos enganosos.

O comportamento coordenado nos comentários do YouTube é muitas vezes impulsionado pelo desejo de fazer determinado conteúdo parecer mais popular ou amplamente apoiado do que realmente é. Esse impulsionamento é normalmente alcançado através de comentários que aumentam artificialmente as métricas de engajamento do vídeo, fazendo com que ganhe mais visibilidade nos algoritmos de recomendação da plataforma. Os atores por trás do CIB buscam influenciar o discurso político, espalhar desinformação ou promover agendas específicas (de-Lima-Santos & Ceron, 2023a).

Tais estratégias de impulsionamento são utilizadas por *social scam bots* (SSB) para promover, além de desinformação, esquemas de fraude ou roubo de informação (Na et al., 2023). SSBs no

YouTube se diferenciam por seus comentários semanticamente relevantes para o contexto do vídeo, impondo desafios à sua detecção. Em alguns casos, SSBs copiam, com leves alterações de palavras e termos, outros comentários de usuários autênticos, alcançando números semelhantes de impulsionamento. Para se manter entre os comentários mais relevantes, SSBs foram observados interagindo com os comentários de outros SSBs. Ao se passarem por perfis autênticos, esses SSBs conseguem atrair usuários para os seus canais, onde links sobre conteúdos apelativos (por exemplo, *vouchers* de jogos ou ganhos financeiros) os direcionam para sites externos (Na et al. 2023).

1.2 Marco teórico

O estudo do CIB cruza várias áreas teóricas importantes, incluindo o conceito de desordem informacional, câmaras de eco, filtros bolha e comunicação política (de-Lima-Santos & Ceron, 2023b). Wardle & Derakhshan (2017) argumentam que a desordem informacional é um fenômeno amplo que se manifesta em três tipos: *misinformation*, *malinformation* e *disinformation* - referindo-se, respectivamente, à disseminação não intencional de informações falsas ou distorcidas, a descontextualização deliberada de informações verdadeiras, e a produção e disseminação intencional de informações falsas ou distorcidas de enganar e prejudicar pessoas, organizações ou grupos sociais. Em geral, o CIB envolve desinformação, uma vez que os atores por trás das campanhas coordenadas têm o objetivo de enganar e manipular a percepção pública (Giglietto et al., 2020b; Mannocci et al, 2024).

Um aspecto importante dessa manipulação é a criação de câmaras de eco (Terren & Borge, 2021). Esse conceito refere-se a situações em que os indivíduos são expostos principalmente a conteúdos que reforçam suas crenças preexistentes, levando a visões mais extremas (Möller, 2021). Nas plataformas como o YouTube, os algoritmos desempenham um papel crucial na formação dessas câmaras ao promover conteúdos com os quais os usuários já interagiram, criando um ciclo de retroalimentação (Möller, 2021). Quando o CIB infiltra-se nessas câmaras, ele amplifica o efeito, dificultando o acesso dos usuários a perspectivas diversas (de-Lima-Santos & Ceron, 2023b).

No contexto do YouTube, o CIB nos comentários pode criar a ilusão de consenso em torno de ideias ou pontos de vista específicos, reforçando o viés de confirmação e a polarização política.

Ao coordenar comentários para apoiar uma narrativa particular, os atores podem manipular a percepção pública sobre um tópico, o que contribui ainda mais para o fenômeno das câmaras de eco online (de-Lima-Santos & Ceron, 2023b).

1.3 Comportamento online e manipulação

A manipulação do comportamento online é uma área crescente de preocupação nas pesquisas acadêmicas, especialmente quando se trata da amplificação de conteúdo por meios inautênticos (de-Lima-Santos & Ceron, 2021a). Em plataformas como o YouTube, a seção de comentários desempenha um papel significativo na maneira como os vídeos são percebidos pelo público em geral. Quando os usuários veem inúmeros comentários positivos ou negativos, são influenciados por eles, independentemente da autenticidade desses comentários (Giglietto et al., 2018). Isso é um exemplo de “prova social” (*social proof* or *informational social influence*, em inglês), um fenômeno psicológico em que as pessoas assumem que as ações de outros refletem o comportamento correto em uma situação dada (Cialdini, 2013).

O CIB aproveita essa tendência ao criar a aparência de apoio ou oposição generalizada a certos conteúdos. Ao coordenar múltiplas contas para inundar a seção de comentários com mensagens falsas ou enganosas, esses atores podem influenciar as opiniões de usuários genuínos e aumentar a visibilidade de certos vídeos (Rocha et al, 2024; Bernardi et al, 2022). Esta tática é frequentemente usada em campanhas políticas, esforços de marketing ou campanhas de desinformação que buscam manipular a opinião pública sobre tópicos controversos, como eleições ou questões de saúde pública (Giglietto et al., 2020b; de-Lima-Santos & Ceron, 2021b).

Além disso, os incentivos econômicos associados ao algoritmo do YouTube também desempenham um papel na prevalência do CIB. Os criadores são recompensados com base em visualizações e engajamento, o que incentiva a produção e promoção de conteúdo que gera controvérsia ou altos níveis de engajamento, independentemente de sua precisão ou autenticidade (McCrosky & Geurkink, 2021). Isso cria um terreno fértil para desinformação, uma vez que os atores podem explorar o sistema para promover sua agenda manipulando métricas de engajamento por meio de CIB.

2. Relevância do projeto para o desenvolvimento científico, tecnológico ou de inovação

Esse projeto de estudo do CIB, particularmente nos comentários do YouTube, oferece um potencial significativo para o avanço tanto da compreensão científica quanto da inovação tecnológica - sobretudo a partir do desenvolvimento de um classificador baseado em IA para a detecção do CIB treinado especificamente para o contexto brasileiro. A relevância deste projeto está em sua nova contribuição para os campos de pesquisa em mídias sociais, comunicação digital e métodos computacionais para detectar e mitigar comportamentos online inautênticos em uma plataforma que ainda não foi explorada e que confia em algoritmos de recomendação influenciados por atitudes dos usuários comentários para oferecer conteúdo que é relevante para eles.

Devido ao número limitado de datasets públicos categorizados e contextualizados para a realidade brasileira, o projeto deve definir protocolos transparentes para a anotação de perfis considerados inautênticos, categorização de vídeos e implementação de modelos. Pesquisas anteriores argumentaram que a transparência em todo o processo de desenvolvimento de modelos é fundamental para minimizar viesamentos sistemáticos, seja pela publicidade de datasets, como pelo questionamento da academia sobre protocolos e métricas de treinamento (Sen et al., 2021; Gallwitz & Kreil, 2021; Hovy & Prabhumoye, 2021).

2.1 Contribuições científicas

Do ponto de vista científico, a investigação do CIB nos comentários do YouTube oferece *insights* sobre como as plataformas de mídias digitais podem ser manipuladas e as implicações mais amplas dessas atitudes para o discurso público. Embora o CIB tenha sido amplamente estudado em plataformas como Facebook e Twitter/X (Giglietto et al., 2020a, 2020b; de-Lima-Santos & Ceron, 2021a, 2021b), o YouTube permanece relativamente inexplorado, especialmente em termos da manipulação que ocorre dentro de suas seções de comentários. Ao focar no YouTube, este projeto contribuirá para a literatura sobre desinformação em plataformas, recomendação algorítmica e o papel das plataformas baseadas em vídeo na formação da opinião pública.

Esta pesquisa também aprimora nossa compreensão das câmaras de eco e de como elas são reforçadas por comportamentos inautênticos. As câmaras de eco, onde os usuários são expostos apenas a conteúdos que alinham-se com suas visões, têm sido um tema de interesse nos estudos de comunicação política e mídia digital. Apesar de opiniões divergentes sobre a existência dessas câmaras de eco (Bruns, 2021), algoritmos de recomendação podem influenciar na sua formação (Möller, 2021). Este projeto contribuirá para o debate contínuo sobre como as plataformas digitais contribuem para a polarização política e a disseminação de desinformação, particularmente mostrando como o comportamento coordenado nos comentários pode exacerbar essas questões.

Outra contribuição importante está na área da transparência algorítmica e moderação de conteúdo. O algoritmo de recomendação do YouTube foi criticado por promover conteúdo controverso (Santini et al., 2023b, McCrosky & Geurkink, 2021), e este projeto pode esclarecer como o engajamento inautêntico na seção de comentários influencia as decisões algorítmicas. Compreender as mecânicas do CIB poderá auxiliar pesquisadores e operadores de plataformas a desenvolver melhores ferramentas para detectar e mitigar seus efeitos

Finalmente, o desenvolvimento de aplicações em redes sociais especializadas para a realidade brasileira é fundamental para diminuir a exposição dos usuários brasileiros de redes sociais a conteúdo nocivo, desinformação e fraude. Apesar da existência de soluções automáticas, quando os dados de um modelo são usados de forma genérica, em aplicações fora do domínio ou em diferentes contextos culturais (González-Bailón & Paltoglou, 2015), os modelos enfrentam desafios fundamentais de precisão devido à falta de flexibilidade para se adaptar a diferentes sinais de coordenação, tipos de campanhas, idiomas e eventos ao longo do tempo.

2.2 Avanços tecnológicos

Do ponto de vista tecnológico, o projeto oferece oportunidades para desenvolver novas ferramentas e métodos de detecção de comportamentos coordenados em mídias sociais. As técnicas atuais para identificar CIB em plataformas de mídias sociais geralmente utilizam análise de redes e detecção de padrões no comportamento das contas (Giglietto et al., 2020a; de-Lima-Santos & Ceron, 2021a). No entanto, o YouTube apresenta desafios únicos devido à natureza de seu conteúdo e das interações dos usuários. Os comentários em vídeos são muitas

vezes menos estruturados do que outras formas de engajamento, o que torna mais difícil aplicar métodos tradicionais de detecção de CIB.

Este projeto tem o potencial de avançar métodos computacionais para detecção de CIB, focando nas características únicas dos comentários no YouTube. Por exemplo, o projeto pode desenvolver algoritmos baseados em técnicas de aprendizado de máquina e processamento de linguagem natural (subcampos da IA) que analisam padrões no tempo de publicação dos comentários, similaridade de conteúdo e comportamento das contas para identificar instâncias de coordenação. Dessa forma, além do conteúdo idêntico, vai permitir identificar postagens similares e alteradas por esses agentes para escapar dos algoritmos de moderação de conteúdo. Além disso, o projeto pode explorar como o conteúdo dos comentários é manipulado para criar determinadas narrativas.

Os dados coletados de comentários serão utilizados para o treinamento de grandes modelos de linguagem (*large language models* – LLMs) específicos para a realidade de parte do YouTube no Brasil. Podemos aprimorar a identificação de padrões culturais e demográficos específicos através da especialização de LLMs já pré-treinados para o português, como BERTimbau (Souza et al., 2020) ou Maritaca (Pires et al., 2023). Posteriormente, esses modelos podem ser continuamente ajustados por novos dados e anotações, bem como disponibilizados para outros colaboradores científicos utilizarem nas suas próprias aplicações. A implementação de modelagens especializadas para a realidade brasileira, além de reforçar a soberania e independência nacional em relação a tecnologias de IA, mitiga efeitos de enviesamento (Hovy & Prabhumoye, 2021).

Esses avanços tecnológicos podem ser aplicados além do YouTube, fornecendo ferramentas para detectar CIB em outras plataformas onde os comentários ou conteúdo gerado pelos usuários desempenham um papel significativo. Além disso, os *insights* obtidos a partir desta pesquisa podem informar o desenvolvimento de sistemas de moderação de conteúdo mais eficazes, que vão além da simples detecção de palavras-chave e incorporam análise comportamental para identificar atividades inautênticas.

2.3 Implicações para a inovação

Os resultados deste projeto podem ter implicações mais amplas para a inovação no campo da governança de mídias sociais e tecnológico. À medida que as plataformas enfrentam o crescente problema da desinformação e do CIB, há uma necessidade clara de soluções brasileiras e inovadoras que equilibrem a proteção da liberdade de expressão com a necessidade de manter a integridade do discurso público no país. Ao fornecer uma compreensão mais profunda de como o CIB opera no YouTube, este projeto poderá informar o desenvolvimento de novas políticas e ferramentas tecnológicas que abordem melhor esses desafios, que sejam treinadas para a língua portuguesa e o contexto brasileiro.

Em particular, o projeto poderá contribuir para o desenvolvimento de ferramentas de moderação de conteúdo mais sofisticadas, que utilizam aprendizado de máquina para detectar comportamento coordenado. Essas ferramentas poderiam ser integradas à plataforma do YouTube para identificar e sinalizar comentários inautênticos, ajudando a reduzir a disseminação de desinformação e teorias conspiratórias. Além disso, a pesquisa poderá informar o design de interfaces de usuário que facilitem a identificação de conteúdo inautêntico, como a inclusão de indicadores de comportamento suspeito nas seções de comentários.

Como forma de contribuição para a comunidade científica, serão desenvolvidos LLMs treinados com comentários no YouTube. Isso poderá ajudar esses modelos a entenderem melhor como é a comunicação brasileira. A experiência em anotações de perfis e treinamento de modelos de ML é fundamental para a identificação de possíveis vieses, aumentando a confiabilidade do LLM especializado. Diversas fontes de vieses foram identificadas na literatura durante o desenvolvimento de modelos de PLN (Hovy, 2021) e esses devem ser entendidos, mitigados e contabilizados (Ntoutsi et al., 2020).

Finalmente, esta pesquisa tem o potencial de influenciar discussões políticas sobre a regulação de plataformas. À medida que governos e órgãos reguladores consideram como enfrentar os desafios impostos pela desinformação e pelo CIB, os resultados deste projeto poderão fornecer evidências valiosas para apoiar o desenvolvimento de regulamentações eficazes. Ao demonstrar como o CIB opera no YouTube e seu impacto no discurso público, o projeto poderá ajudar a moldar futuros marcos regulatórios que protejam os usuários contra manipulações, preservando ao mesmo tempo a natureza aberta das plataformas online.

3. Experiência do coordenador e da equipe

A coordenadora do projeto, Dra. R. Marie Santini, é professora da Escola de Comunicação da Universidade Federal do Rio de Janeiro (UFRJ), atuando nos cursos de graduação, mestrado e doutorado. Pesquisadora-membro da rede europeia VOX-Pol Network of Excellence, financiada pelo Programa da União Europeia 7 (FP7) focado em pesquisar funções e impactos do Extremismo Político Online Violento. Membro do International Observatory on Information and Democracy (OID) fundado por Angel Gurria e Shoshana Zuboff. Membro do Comitê Científico do IPIE (International Panel of Information Environment) fundado pelo professor Philip Howard da Universidade de Oxford, e Membro do Comitê de especialistas do Nobel Prize Summit 2023. Possui experiência de pesquisa nas áreas de Comunicação e Informação com ênfase em Estudos de Internet e Redes Sociais. Atua principalmente nos seguintes temas: curadoria algorítmica e sistemas de recomendação online; manipulação e desinformação nas redes sociais; social bots; propaganda computacional; comportamento e manipulação da opinião pública. Autora dos livros “O Algoritmo do Gosto” – volume 1 e 2 – publicados em 2020, além de dezenas de artigos científicos em revistas nacionais e internacionais. Lidera como Investigadora Principal os projetos de pesquisa “Máquinas de opinião: propaganda computacional, contágio e desinformação nas redes sociais online” e “The media ecosystem dynamic regarding environmental issues, land use, indigenous people's rights and support for climate policies in Brazil” e “Propaganda Computacional, Desinformação e a Indústria da Influência Digital”. É fundadora e diretora do NetLab UFRJ - Laboratório de Estudos de Internet e Redes Sociais (<https://netlab.eco.ufrj.br/>), onde desenvolve pesquisa, ensino e atividades de extensão entre a universidade e a sociedade civil no Brasil e no mundo.

Composto por mais de 40 pesquisadores, o NetLab UFRJ atua de maneira plural e com equipe interdisciplinar para desenvolver sua agenda de pesquisas de impacto, com amplos esforços de divulgação científica, participação constante no debate público e engajamento em projetos em parceria e cooperação técnica com diversas organizações no âmbito da proteção das instituições democráticas e da garantia de direitos à comunicação, à liberdade de expressão, e à integridade da informação. Sob direção da Dra. R. Marie Santini, o NetLab UFRJ integra iniciativas e grupos de trabalho internacionais, como o International Panel on the Information Environment (IPIE), a rede de pesquisa em extremismo e terrorismo online VOX-Pol, a Sala de Articulação contra a Desinformação (SAD), e o International Observatory on Information &

Democracy, que agrega pesquisas da área e oferece aos formuladores de políticas públicas avaliações globais periódicas sobre informação e comunicação, e seus impactos na democracia.

As atividades da Dra. R. Marie Santini junto ao NetLab UFRJ têm como um dos objetivos principais a publicação de artigos em periódicos científicos internacionais e nacionais de alto impacto e prestígio nas áreas da comunicação, computação e ciência da informação, bem como o diálogo com os pares por meio da participação em congressos e eventos científicos. Entre os congressos e conferências, destacamos as participações no Cambridge Disinformation Summit, Encontro Anual da Compós, Terrorism and Social Media Conference (TASM), ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT) e no Encontro Nacional de Pesquisa em Ciência da Informação (ENANCIB). As pesquisas da Dra. R. Marie Santini no NetLab UFRJ foram publicadas em artigos revisados por pares em mais de 30 revistas e anais de congressos, entre elas Policy & Internet, Social Media + Society, International Review of Information Ethics, Communication Studies, Journal of Communication, Information and Ethics in Society, Lumina, Liinc em Revista, International Journal of Communication, Convergence: The International Journal of Research into New Media Technologies, Compós, Comunicação & Sociedade e Mídia e Cotidiano. A essa lista somam-se mais de 45 relatórios técnicos e notas técnicas que têm contribuído com o campo de estudos sobre internet e plataformas digitais, e gerado impactos na sociedade civil e na elaboração de políticas públicas, alimentando o debate na imprensa.

Desde 2023, o NetLab UFRJ teve mais de 400 aparições na mídia, tanto online quanto impressa e televisiva (NetLab UFRJ, [s.d.]), com uma média de aproximadamente cinco aparições por semana. Destacam-se as contribuições com veículos de amplo alcance nacional, como o Jornal Nacional, UOL, R7, Folha de S. Paulo, Jornal Hoje, Jornal da Globo, Fantástico, Valor Econômico, Estadão, SBT News, Record News, CNN Brasil, G1, O Globo, Jornal da Band, Le Monde Diplomatique, Metrôpoles, entre outros. As pesquisas também servem para embasar materiais para a mídia especializada de alto nível, como Agência Pública, Piauí, The Intercept, Núcleo, Nexo, Desinformante etc. Na mídia internacional, os trabalhos do NetLab já repercutiram em veículos como The Guardian, The Washington Post, Associated Press e La Nación.

Em 2023, a Dra. R. Marie Santini foi a única representante da América Latina no comitê consultivo da Cúpula do Prêmio Nobel 2023. Organizado em parceria com o Programa das Nações Unidas para o Desenvolvimento (PNUD) e a Aliança de Bens Públicos Digitais (DPGA –

Digital Public Goods Alliance), a articulação tem como objetivo identificar e promover soluções digitais de código aberto que possam ser ampliadas para tratar do problema da desinformação em suas raízes. Já em 2024, apresentou um relatório sobre mudança climática (NetLab UFRJ, 2024c) no Grupo de Trabalho de Economia Digital do G20, em diálogo com organizações interessadas na pauta ambiental e publicou, em agosto, um *policy brief* no G20 com recomendações sobre mecanismos de transparência, parâmetros e meios de acesso a dados de plataformas digitais (Santini et al., 2024a).

O parceiro deste projeto, **Dr. Mathias Felipe de Lima Santos**, é um especialista de destaque em mídia e ciência da computação, com um histórico reconhecido internacionalmente no estudo de tecnologias aplicadas ao consumo de informação, como desinformação, jornalismo de dados, automação e inteligência artificial. Atualmente, atua como professor na Macquarie University (Austrália). Embora tenha concluído o doutorado apenas em outubro de 2021, o coordenador já editou três livros (sendo dois focados em Fact-checking no Sul Global a serem publicados no próximo ano pela Palgrave MacMillan) e mais de 35 publicações, incluindo artigos revisados por pares e capítulos de livros. De acordo com a base de dados SciVal, está em 1º lugar em produtividade na temática inteligência artificial (IA), big data e jornalismo (Tópico: T.20729) entre 2019 e 2024.

Mathias Felipe já foi pesquisador de pós-doutorado na Universidade de Amsterdã (UvA), onde trabalhou no projeto Human(e) AI, focando em pesquisa sobre IA responsável e o papel do viés em algoritmos e IA na indústria de mídia. Além disso, contribuiu com a equipe do consórcio AI4Media, parte do programa de pesquisa e inovação Horizonte 2020 da União Europeia. Neste projeto, atuou em um consórcio que inclui mais de 30 parceiros nas áreas de tecnologia e mídia, entre universidades, centros de pesquisa e parceiros da indústria.

O coordenador também está engajado em uma colaboração global com acadêmicos de diferentes países, estudando o crescente poder econômico e influência política das plataformas sobre as inovações tecnológicas na indústria de informação, com foco especial na Iniciativa Google News. Liderando este esforço, que inclui acadêmicos das Américas, África e Ásia. Neste projeto, ele investiga as implicações estruturais das tecnologias emergentes, particularmente a IA, dentro das redações de mídia, e como essas tecnologias podem reforçar as dependências existentes dessas organizações em relação à indústria de tecnologia, afetando simultaneamente o público no Sul Global. Através desse projeto, tem ganhando grande alcance

mediático, aparecendo em veículos de comunicação como *The Conversation*, *NiemanLab*, *Nation Africa*, *Poder360*, *LatAm Journalism Review*, entre outros.

Desde 2021, está como pesquisador associado na Universidade Federal de São Paulo (Unifesp), onde desempenhou um papel crucial na criação do Observatório de Mídia Digital e Sociedade (DMSO). Neste ambiente, supervisiona estudantes de graduação e mestrado em diversos projetos de pesquisa, incluindo desinformação política visual e estudos de plataformas cruzadas. No DMSO, publicou pesquisas no campo de comportamento de desinformação e comportamento coordenado em livros de grandes editoras (de-Lima-Santos & Ceron, 2023a; 2023b) e jornais com um alto fator de impacto, como *Online Social Networks & Media* and *Social Network Analysis & Mining* (Ceron et al., 2021a; 2021b).

Mathias Felipe também foi beneficiário de uma prestigiosa bolsa Marie Skłodowska-Curie para jovens pesquisadores no projeto JOLT, parte do programa Horizon 2020, que financiou seu doutorado em Comunicação na Universidade de Navarra, Espanha. Durante este projeto, desenvolveu um conhecimento profundo sobre a interseção entre mídia, tecnologia e negócios, especialmente em relação às mudanças trazidas pelo uso de tecnologia na produção de notícias. Trabalhou diretamente com importantes organizações de mídia, como Al Jazeera, BBC, La Nación e ProPublica, explorando a aplicação de jornalismo de dados, automação e inteligência artificial.

Além disso, cursou o prestigiado mestrado Erasmus+ em Digital Communication Leadership (DCLead) oferecido pela Universidade de Aalborg (Dinamarca) e Universidade de Salzburgo (Áustria). Através desse programa, passou um semestre em 2017 na Queensland University of Technology (QUT), Austrália, onde conduziu pesquisas extensivas sobre o uso de jornalismo de dados na indústria de notícias australiana. Trabalhou com professores renomados no campo de redes sociais e desinformação, como Axel Bruns e Dr. Aljoshia Karim Schapals, no Digital Media Research Centre (DMRC), classificado entre os melhores centros de pesquisa em mídia digital na Austrália.

Mathias Felipe é convidado a contribuir para volumes editados de prestígio internacional e diversas revistas acadêmicas, publicando com editoras reconhecidas, como Springer, Palgrave, Routledge e MDPI Books, destacando sua ampla rede de colaborações e sua relevância no campo da pesquisa em mídia e desinformação.

Dr. Francisco Q. Camargo (Chico Camargo) também fará parte da equipe que lidera esse projeto. Ele é professor no Departamento de Ciência da Computação da Universidade de Exeter, pesquisador associado no Oxford Internet Institute, da Universidade de Oxford, membro do conselho da Sociedade de Sistemas Complexos, Líder do Tema de Ciências Sociais Computacionais no Instituto de Ciência de Dados e Inteligência Artificial da Universidade de Exeter, Diretor Adjunto no Centro de Comunicação sobre Clima e Ciência de Dados (C3DS), professor visitante no Departamento de Língua e Literatura Inglesa da Ewha Womans University, na Coreia do Sul, Fellow do Alan Turing Institute e diretor do CC Lab. Sua pesquisa consiste em desenvolver e aplicar novos métodos computacionais para as ciências sociais, frequentemente envolvendo a análise em larga escala de conteúdo de mídia tradicional e de redes sociais. Seu trabalho usa diversas ferramentas de aprendizado de máquina, processamento de linguagem natural, análise de redes sociais, modelagem matemática de sistemas complexos e computação de alto desempenho, além de teorias sobre comportamento humano, cultura e sociedade. Seu trabalho multidisciplinar com físicos, psicólogos, matemáticos, antropólogos, jornalistas e cientistas políticos já resultou em múltiplas oportunidades de colaboração com órgãos públicos, e já foi financiado pelo Alan Turing Institute, DSO Singapore National Laboratories, Meta Foundational Integrity & Social Impact Research, Royal Institution e Mozilla Foundation, além do Children's Investment Fund.

O professor Chico também tem experiência como divulgador científico, já tendo escrito para a Science, HuffPost Brasil, The Conversation, e produzido mais de 50 vídeos para o YouTube. Ele é atualmente um dos diretores do Science Vlogs Brasil.

Antes de ingressar na Universidade de Exeter, Chico foi um pesquisador de pós-doutorado em ciência de dados no Oxford Internet Institute, na Universidade de Oxford, onde trabalhou com o Prof. Scott Hale estudando a dinâmica da opinião pública, volatilidade política, definição de agendas políticas (agenda-setting) e mobilidade humana. Antes disso, fez seu D.Phil. (PhD) em Biologia de Sistemas, também em Oxford, como Clarendon Scholar no Brasenose College. Em sua pesquisa de doutorado, trabalhando com Ard Louis, ele usou ferramentas de sistemas complexos e aprendizado de máquina para investigar os princípios físicos que regem a biologia evolutiva. Durante seu PhD, também trabalhou com Kevin Foster e William (Mack) Durham desenvolvendo uma pipeline computacional para analisar comportamento coletivo em bactérias. Além disso, foi um visitante acadêmico no Wolfson Centre for Mathematical Biology, trabalhando com Philip Maini e Ruth Baker. Antes de se mudar para o Reino Unido, Francisco formou-se bacharel pela Universidade de São Paulo, como parte do Curso de Ciências

Moleculares, e trabalhou em modelagem matemática de processos biológicos no Instituto de Física Teórica da UNESP, colaborando com Roberto Kraenkel e Frederico Gueiros Filho.

4. Objetivos gerais e específicos

Objetivo geral:

O objetivo geral deste projeto é investigar e desenvolver uma compreensão aprofundada do CIB nas seções de comentários do YouTube, com foco na identificação de padrões de manipulação, nas formas de amplificação de conteúdo e nas suas implicações para a opinião pública e os algoritmos de recomendação da plataforma, especialmente no contexto brasileiro. Através desse estudo, nosso plano é desenvolver uma ferramenta que permita que pesquisadores, jornalistas e atores da sociedade civil possam identificar automaticamente esses casos de comportamento coordenado tentando mitigar o impacto desses conteúdos.

Objetivos específicos:

1. **Identificar padrões de comportamento coordenados** nas seções de comentários do YouTube, diferenciando contas inautênticas (bots e contas falsas) de contas genuínas.
2. **Desenvolver uma ferramenta que utiliza técnicas baseadas em aprendizado de máquina e processamento de linguagem natural** para detectar automaticamente CIB nos comentários.
3. **Analisar o impacto do CIB nos algoritmos de recomendação do YouTube**, determinando como o engajamento artificial afeta a visibilidade de conteúdos específicos.
4. **Mapear narrativas e temas prioritários** que são amplificados através do CIB no YouTube, com ênfase em tópicos de relevância pública como política, saúde e desinformação.
5. **Fornecer recomendações para mitigação e moderação de CIB** no YouTube, colaborando com plataformas e reguladores para fortalecer a integridade do discurso público online.

5. Metas e indicadores da proposta

Com base nos objetivos específicos descritos acima, foram definidas as seguintes metas e indicadores:

- **Meta 1: Identificação de padrões de CIB nas seções de comentários do YouTube**
Indicador: Criação de um banco de dados categorizado com pelo menos 100 vídeos com potencial para CIB ao longo dos primeiros 6 meses.
- **Meta 2: Desenvolvimento de algoritmos de ML e PLN para a detecção de CIB**
Indicador: Implementação de um protótipo funcional que alcance uma precisão de pelo menos 85% na detecção de CIB em comentários automatizados, dentro dos primeiros 12 meses.
- **Meta 3: Análise do impacto do CIB nos algoritmos de recomendação**
Indicador: Publicação de um relatório técnico que demonstre como o engajamento artificial afeta a recomendação de vídeos, com base em experimentos conduzidos na plataforma.
- **Meta 4: Publicação de relatórios e artigos acadêmicos**
Indicador: Publicação de pelo menos três artigos acadêmicos em periódicos relevantes nas áreas de comunicação, inteligência artificial e estudos de desinformação.
- **Meta 5: Recomendação de políticas de mitigação para plataformas**
Indicador: Apresentação de um conjunto de recomendações práticas para o YouTube e órgãos reguladores, com base nos achados do projeto, ao final de 24 meses.

6. Metodologia

A metodologia deste projeto será estruturada em cinco etapas principais, todas interligadas e voltadas para a análise detalhada do CIB nas seções de comentários do YouTube.

Etapas 1: Coleta de dados

A primeira etapa envolve a coleta de dados de comentários do YouTube, focado em vídeos sobre tópicos de alto impacto, como política, saúde pública e desinformação. Para essa coleta, utilizaremos a API pública do YouTube (YouTube Data API) para extrair comentários de vídeos populares, buscando especialmente aqueles relacionados a eventos políticos (como eleições) e temas de crises de saúde. Além disso, a coleta incluirá dados temporais (frequência de

comentários e curtidas) para auxiliar na identificação de possíveis comportamentos inautênticos coordenados.

O foco será dado a vídeos em português, especialmente em canais influentes no Brasil, garantindo que o contexto local seja refletido nos resultados. A expectativa é coletar e identificar pelo menos 100 vídeos para que alcancemos 100.000 comentários ao longo dos primeiros seis meses do projeto.

Etapa 2: Processamento e limpeza dos dados

Após a coleta, será necessário realizar um processamento e limpeza dos dados para remover comentários irrelevantes, spam e duplicatas. Técnicas de pré-processamento, como tokenização, remoção de stopwords e *stemming*, serão aplicadas para preparar os dados para análise mais aprofundada.

Além disso, será realizada uma filtragem para identificar padrões de repetição, como comentários muito semelhantes postados em diferentes vídeos ou ao longo de curtos intervalos de tempo, o que pode indicar a presença de redes automatizadas de comentários.

Etapa 3: Detecção de CIB usando aprendizado de máquina e PLN

Nesta fase, desenvolveremos um sistema de detecção de CIB em comentários do YouTube utilizando algoritmos de ML e PLN para detectar CIB nos dados coletados. Usaremos modelos supervisionados e não supervisionados para identificar padrões anômalos, como:

- **Clusterização:** Modelos de clusterização não supervisionados, como *K-means*, serão utilizados para agrupar comentários semelhantes, buscando padrões que possam indicar coordenação.
- **Classificação:** Modelos supervisionados, como *Random Forest* e *Support Vector Machines (SVM)*, serão treinados com dados rotulados (comentários identificados como inautênticos) para identificar novos casos de CIB.
- **PLN:** Modelos de PLN, como BERT ou transformers, serão usados para analisar o conteúdo textual dos comentários, identificando similaridades semânticas e padrões linguísticos que podem sugerir coordenação entre contas.

O objetivo é desenvolver um modelo preditivo com precisão superior a 85% na detecção de comportamentos inautênticos.

Etapa 4: Análise do Impacto nos Algoritmos de Recomendação

Além da identificação do CIB, investigaremos como o engajamento artificial gerado por comentários coordenados afeta o algoritmo de recomendação do YouTube. Para isso, analisaremos as sugestões de recomendação do YouTube para verificar a existência de comentários similares nos diferentes vídeos sugeridos pela plataforma. Dessa forma, poderemos mostrar construir redes que mostram as relações entre esses diferentes vídeos que foram artificialmente providos através dos comentários.

Etapa 5: Divulgação de Resultados e Recomendação de Políticas

Finalmente, consolidaremos os resultados das fases anteriores em relatórios técnicos e artigos acadêmicos. Também elaboraremos um conjunto de recomendações práticas para o YouTube e outros stakeholders, propondo estratégias de mitigação para reduzir a prevalência do CIB na plataforma.

Além de publicações acadêmicas, o projeto se compromete a dialogar com reguladores brasileiros e internacionais sobre os impactos do CIB, colaborando para o desenvolvimento de políticas públicas e diretrizes para as plataformas digitais.

7. Cronograma de execução do projeto

O cronograma foi projetado para garantir que todas as etapas sejam executadas de maneira integrada, com espaço suficiente para ajustes metodológicos e análise dos resultados ao longo dos 24 meses, como descrito pelo CNPq.

| Etapa | Descrição | Mês de Início | Previsão de duração | Mês de Conclusão |
|-------|--|---------------|---------------------|------------------|
| 1 | Implementação da infraestrutura de pesquisa. Definição de parâmetros e metodologia para coleta de dados. Início da coleta de dados preliminar. | 1º | 3 meses | 3º |
| 2 | Coleta intensiva de dados de comentários no YouTube. Limpeza e pré-processamento de dados. | 4º | 3 meses | 6º |
| 3 | Implementação de algoritmos de ML e PLN para detecção de CIB. Treinamento inicial. | 7º | 3 meses | 9º |
| 4 | Refinamento dos modelos de detecção de CIB. Coleta contínua de novos dados. Análises preliminares. | 10º | 3 meses | 12º |
| 5 | Análise do impacto do CIB nos algoritmos de recomendação do YouTube. Experimentos controlados. | 13º | 3 meses | 15º |
| 6 | Desenvolvimento do protótipo final de detecção de CIB. Publicação de artigos intermediários. | 16º | 3 meses | 18º |
| 7 | Revisão de dados e aprimoramento das técnicas de detecção. Início da elaboração das recomendações. | 19º | 3 meses | 21º |
| 8 | Finalização de publicações. Apresentação de resultados a diferentes stakeholders. | 22º | 3 meses | 24º |

8. Aspectos Éticos e Legais

Este projeto seguirá rigorosamente as diretrizes éticas e legais aplicáveis à pesquisa com seres humanos e dados sensíveis:

1. **Consentimento informado:** Será garantido que todos os participantes, caso ocorram entrevistas e/ou pesquisas, forneçam consentimento informado, estando cientes dos objetivos, métodos, potenciais benefícios e riscos associados ao estudo.
2. **Privacidade e confidencialidade:** Medidas serão implementadas para proteger a privacidade e a confidencialidade dos dados coletados. Está previsto apenas a coleta de dados públicos. Dados pessoais, caso sejam coletados, serão anonimizados e armazenados de forma segura, acessíveis apenas aos pesquisadores autorizados. A equipe será treinada em práticas de manuseio seguro de dados, e a conformidade com regulamentações de proteção de dados, como a LGPD (Lei Geral de Proteção de Dados), caso seja necessário.

3. **Revisão por comitês de ética:** O projeto será submetido à revisão de comitês de ética de instituições de pesquisa envolvidas, garantindo a conformidade com normas nacionais e internacionais para a pesquisa ética. Ajustes serão feitos conforme necessário para atender a todas as recomendações e exigências regulatórias.
4. **Uso responsável de tecnologia:** O uso de LLMs e outras tecnologias de IA será conduzido com atenção às questões de viés algorítmico e impacto social, evitando a disseminação de informações errôneas e garantindo a transparência nos processos de tomada de decisão automatizada. Todo conteúdo produzido será devidamente etiquetado como produzido por IA.

9. Disseminação dos resultados esperados

A tradução e disseminação eficaz dos resultados do projeto são fundamentais para maximizar seu impacto e aplicabilidade. Os resultados serão disseminados através de um plano abrangente de tradução e comunicação científica. Isso incluirá a publicação de artigos em revistas acadêmicas, desenvolvimento de relatórios de políticas públicas, organização de workshops/seminários, e eventos com *stakeholders*. Esses são detalhados abaixo:

Publicações acadêmicas e relatórios: Serão elaborados artigos científicos para publicação em revistas de alto impacto e relatórios técnicos direcionados a formuladores de políticas públicas, detalhando os achados do projeto e recomendações práticas.

Engajamento com mídia digital: A colaboração com mídia nacional e internacional também será explorada para amplificar o alcance dos resultados e promover um diálogo contínuo sobre CIB no YouTube.

Participações em conferências acadêmicas: Os resultados serão disseminados em eventos acadêmicos de alcance internacional tanto no campo da comunicação quanto da ciência da computação, como ICA (*International Communication Association*), IAMCR (*International Association for Media and Communication Association*), AoIR (*Association of Internet Researchers*), ASONAM (*Advances in Social Network Analysis and Mining*), and IEEE Network Science and Engineering.

Participações em eventos da indústria: Os resultados serão disseminados em eventos com grande participação de stakeholders da indústria de mídia como o Global Fact, que será realizado no ano de 2025 no Rio de Janeiro.

10. Orçamento

Recursos solicitados - Custeio e capital

Custeio

| Itens de dispêndio | Valor total estimado | Detalhamento | Justificativa |
|--------------------------------------|----------------------|--|--|
| Despesas acessórias com importação | R\$20.000,00 | Frete, seguro e desembaraço na alfândega para compra dos dois computadores Dell indicados nas despesas de capital. | Os materiais de TI listados nas despesas de capital serão utilizados para o cluster de computadores do projeto. Esta aquisição irá expandir consideravelmente a capacidade computacional do laboratório, fomentando pesquisas futuras. Os equipamentos adquiridos serão de uso exclusivo para pesquisas científicas. Não há empresas nacionais que produzam estes bens, ou similares, no país. |
| Diárias (total) | R\$ 56.460,00 | Trinta diárias no país de R\$320 para os dois pesquisadores brasileiros no exterior (2 viagens para dois pesquisadores por ano, de duração de sete dias); dezesseis diárias para viagens internacionais de U\$370; e dez diárias de U\$260 para os coordenadores brasileiros envolvidos (no Brasil e no exterior), considerando que os países têm valores de diárias diferentes. | Estadia dos pesquisadores durante as viagens para reuniões de pesquisa e summer school no Brasil, e para apresentação dos resultados da pesquisa em eventos internacionais listados no projeto (se aprovados os trabalhos), ou em outros eventos internacionais nos quais os artigos forem aprovados. |
| Passagens (total) | R\$ 63.000,00 | Duas passagens Londres-Rio de Janeiro para o pesquisador brasileiro localizado na Inglaterra; duas passagens Sydney-Rio de Janeiro para o pesquisador brasileiro localizado na Austrália; Duas passagens para os EUA e duas para a Europa para apresentação de trabalhos em congressos internacionais | Deslocamentos para reuniões de pesquisa e summer school no Brasil, e para apresentação dos resultados da pesquisa em eventos internacionais listados no projeto (se aprovados os trabalhos), ou em outros eventos internacionais nos quais os artigos forem aprovados. |
| Total de Passagens e Diárias: | R\$ 119.990,00 | | |
| Total de Custeio | R\$ 139.960,00 | | |

Capital

| Itens de dispêndio | Valor total estimado | Detalhamento (pt / en) | Justificativa (pt / en) |
|------------------------------------|-----------------------|---|---|
| Equipamentos e Material permanente | R\$ 456.000,00 | 2x Dell PowerEdge 2U (2x EPYC 9554, 512GB RAM, 480GB SSD) 4x 20TB HDD SAS 1x Switch 24 portas PoE - 4x SFP+ | Aquisição de recursos computacionais para armazenamento e desenvolvimento dos modelos de IA, LLMs e coleta dos dados necessários ao projeto. Esses recursos serão utilizados para criar o cluster computacional e estruturar o datacenter do NetLab. Os custos são uma estimativa baseada no orçamento de \$35.000 por Dell PowerEdge, convertidos a um câmbio de R\$5.60, R\$13.000 por HDD e R\$12.000 para o Switch. |
| Material Bibliográfico | R\$ 4.040,00 | Compra de livros nacionais e importados e acesso a artigos de revistas internacionais que não estejam no Portal da Capes de periódicos. | As publicações originadas deste projeto serão prioritariamente publicadas para acesso livre. Porém necessitaremos de compra de livros e acesso a artigos internacionais pagos, fundamentais para atualização sobre como a literatura científica recente discute a ocorrência de CID nas plataformas e realidades de outros países. |
| Total Capital: | R\$ 460.040,00 | | |
| Total do Projeto: | R\$ 600.000,00 | | |

11. Disponibilidade de infraestrutura física e capacidade técnica da equipe de pesquisa

Sob a direção da Dra. R. Marie Santini, o NetLab UFRJ está sediado na Escola de Comunicação da Universidade Federal do Rio de Janeiro (ECO/UFRJ) e será responsável por parte relevante da execução do projeto. O laboratório conta com uma infraestrutura robusta para realização de atividades de pesquisa, documentação, coleta, análise, processamento e armazenamento de grande quantidade e variedade de dados (big data). Isso inclui uma estrutura física própria, com sala na Escola de Comunicação, além de uma estrutura virtual, formada por servidores de nuvem para coleta e armazenamento de dados e softwares acessíveis online para a colaboração remota.

Porém, para executar esse projeto e futuras pesquisas que envolvam o desenvolvimento de Inteligência Artificial, o NetLab UFRJ precisa construir um cluster de computadores para expandir consideravelmente a capacidade computacional do laboratório e garantir a continuidade das pesquisas no médio e longo prazo. A infraestrutura será necessária para coleta e processamento de dados de diferentes sites e plataformas digitais, capaz de monitorar ininterruptamente conteúdos e fontes definidas em função das agendas de pesquisa definidas. Ressaltamos o caráter único desta empreitada, que não funciona sob a contratação de serviços externos, sendo desenvolvida e mantida inteiramente pela pesquisa do laboratório para atender às suas demandas de pesquisa de forma inteiramente customizada. A partir da infraestrutura que o laboratório vem construindo nos últimos 5 anos, foram coletados dados de mais de 2 milhões de vídeos do YouTube, além de 1 bilhão de publicações no X/Twitter, 70 milhões de mensagens enviadas em grupos públicos do WhatsApp e do Telegram, 1 milhão de anúncios da Meta e 300 mil vídeos do TikTok, e também 8 milhões de publicações em sites noticiosos e cerca de 150 mil anúncios nativos veiculados nestes sites.

Além da infraestrutura do NetLab, nosso laboratório possui acordo de cooperação com o CBPF (Centro Brasileiro de Pesquisas Físicas) para alocação de recursos computacionais na sua estrutura de Data Center. Portanto, o nosso projeto prevê a aquisição e alocação de máquinas de alto desempenho dentro desse Data Center para o desenvolvimento dos modelos de IA. Com isso, o Netlab poderá se beneficiar da infraestrutura de rede, eletricidade e refrigeração do CBPF que possui Data Center multiusuário muito bem estruturado, incrementando a capacidade computacional do Instituto de Pesquisa.

Os pesquisadores possuem formação e experiência em diferentes áreas, como Ciência da Informação, Comunicação, Sociologia, Engenharia, Ciência de Dados e Computação. A expertise técnica e computacional da equipe, bem como sua capacidade de desenvolver novas e inovadoras ferramentas de pesquisa, é reforçada sobretudo pelas publicações científicas, inclusive no âmbito da identificação de comportamento inautêntico em mídias digitais. Um exemplo nesse sentido é o artigo da Dra. R. Marie Santini publicado em pré-print que apresenta o Gotcha, um classificador desenvolvido e otimizado para a detecção de perfis no X/Twitter que publicam em português com comportamento inautêntico coordenado, chamados nesse contexto de social bots, a partir do processamento computacional de centenas de características das publicações e dos perfis em si (Santini et al., 2023a). Esse artigo sobre o classificador seguiu anos de discussões a respeito do comportamento inautêntico coordenado em plataformas de redes sociais e análises sobre como os social bots poderiam servir à

manipulação da opinião pública no Brasil (Santini et al., 2020; Santini et al., 2021b). O Gotcha foi utilizado em uma recente publicação do *The International Review of Information Ethics*, em que pesquisadores do laboratório demonstram como contas inautênticas foram utilizadas para reverberar e ecoar determinados posicionamentos ideológicos sobre os incêndios na Floresta Amazônica em 2020 e inflar o debate de maneira artificial (Salles et al., 2024).

Outros trabalhos envolvem diretamente análises computacionais sobre dados do YouTube, incluindo investigações sobre padrões de comportamento nos comentários e o impacto da desinformação sobre os algoritmos de recomendação da plataforma. Ao realizar modelagem de tópicos em comentários e análise de rede de vídeos recomendados no YouTube, investigamos como as teorias da conspiração dependem de affordances da plataforma e de um cenário político favorável para ganhar adesão e relevância social (Salles et al., 2023). Também utilizamos análises de tópicos para identificar padrões de conteúdo e temporalidade dos comentários em um conjunto de vídeos recomendados na busca por palavras chave, revelando o impacto de contextos políticos no volume de comentários desinformativos em vídeos atuais e antigos (Barros et al., 2024). Para entender como o algoritmo de recomendação do YouTube pode ser guiado por acordos comerciais obscuros aos usuários, realizamos um experimento em que simulávamos o comportamento de novos usuários sem dados históricos e documentamos grupos de mídia sugeridos na página inicial da plataforma (Santini et al., 2023b).

Um trabalho recente do NetLab UFRJ ganhou o prêmio de “Melhor Artigo Curto” no Workshop Brasileiro de Análise de Redes Sociais e Mineração da 44^a edição do Congresso da Sociedade Brasileira de Computação (Ciodaro et al., 2024). Neste artigo, apresentamos uma nova técnica para aplicação de Self-organizing maps para a pesquisa em mídias sociais como uma alternativa às análises de redes tradicionais, que comumente atravessam problemas para lidar com conjuntos de dados massivos.

É importante ressaltar que, diante de tentativas de regulamentação em diversos países nos últimos anos, diferentes empresas de mídias digitais têm agido para dificultar o acesso a dados e armazenamento de dados. A decisão de fornecer ou remover mecanismos de transparência é motivada por fatores comerciais e políticos das plataformas, que, muitas vezes, adotam medidas que impedem a plena auditabilidade do conteúdo veiculado (Bossetta, 2020). Deste modo, as metodologias de coleta de dados descritas acima poderão ser adaptadas em função das limitações impostas pelas plataformas ao longo do período de realização do projeto.

Por isso é fundamental a construção e o investimento em pesquisas que contem com a autonomia de Data Centers em Institutos de Pesquisa e Universidades Públicas, a fim de, a fim de garantir o avanço e a soberania das pesquisas brasileiras que busquem soluções inovadoras para problemas de nosso país, com tecnologia nacional e de ponta.

12. Contrapartida

A **Universidade Federal do Rio de Janeiro (UFRJ)** disponibilizará a infraestrutura necessária para a execução deste projeto. Contudo, para o bom desenvolvimento deste projeto, a infraestrutura existente deverá ser complementada com os equipamentos específicos solicitados, que serão essenciais para a realização de análises avançadas e para o desenvolvimento das tecnologias propostas. A estrutura física do NetLab UFRJ, localizado no campus da Praia Vermelha, é mantida pelo laboratório com verbas de financiamento institucional oriunda de outros projetos, e pela própria universidade, responsável pelo pagamento de despesas de eletricidade e internet e pela limpeza do espaço. Similarmente, todos os pagamentos por equipamentos e serviços oferecidos por pessoas jurídicas, bem como os custos de materiais de consumo, serão cobertos pelas verbas de financiamento institucional do laboratório.

Além disso, a UFRJ, sendo uma instituição amplamente reconhecida na área de mídia e desinformação, facilitará parcerias e colaborações com especialistas renomados do setor. Essa rede de colaboração contribuirá para enriquecer o projeto, proporcionando acesso a conhecimentos especializados e fortalecendo o desenvolvimento de estratégias eficazes para a gestão da desinformação em saúde. Dessa forma, a contrapartida institucional da UFRJ será um elemento chave para o sucesso e o impacto deste projeto.

A **Macquarie University**, na Austrália, também oferecerá uma significativa contrapartida para o projeto proposto, demonstrando seu comprometimento com a pesquisa de alta qualidade. Esta contrapartida inclui tanto recursos humanos quanto tecnológicos, além de infraestrutura e serviços de suporte essenciais. No que diz respeito aos recursos humanos, o Dr. Mathias Felipe de Lima Santos, um pesquisador renomado da instituição, dedicará 10% de seu tempo ao projeto. Esta contribuição é valiosa não apenas em termos financeiros, mas também pelo conhecimento especializado e experiência que ele traz para o projeto.

A universidade também disponibilizará recursos tecnológicos de ponta para o desenvolvimento do projeto. Isto inclui acesso a licenças para software especializado de análise estatística e modelagem, importantes para o desenvolvimento do projeto. O acesso irrestrito à extensa biblioteca digital da Macquarie University, com suas vastas bases de dados e periódicos científicos, também faz parte desta contrapartida. Além disso, a universidade mantém acordos com várias editoras com periódicos de alto fator de impacto para publicação em acesso aberto. Isso é essencial para a disseminação dos resultados da pesquisa, além de alcançar uma maior audiência.

Em termos de infraestrutura, a Macquarie University proporcionará espaço de escritório dedicado no campus para o pesquisador, bem como acesso a salas de reunião para colaborações e apresentações. Estes espaços físicos são essenciais para facilitar o trabalho em equipe e a disseminação dos resultados da pesquisa.

Para garantir o bom andamento do projeto, a universidade oferecerá serviços de suporte abrangentes. Isso inclui suporte administrativo para assistência em gestão de projetos e elaboração de relatórios, além de suporte de TI para manutenção de equipamentos e assistência técnica contínua. Este pacote abrangente de apoio demonstra o compromisso da Macquarie University em fornecer um ambiente propício para pesquisas inovadoras e de alto impacto.

A **University of Exeter**, na Inglaterra, se compromete a fornecer suporte ao projeto, oferecendo tempo de dedicação do pesquisador, infraestrutura humana e tecnológica, além de acesso a recursos online. No que tange aos recursos humanos, o Dr. Chico Camargo, Senior Lecturer em Ciência da Computação na University of Exeter, dedicará 10% de seu tempo ao projeto. Essa contribuição é significativa não apenas em termos financeiros, mas também pelo conhecimento especializado e pela experiência que ele aporta à iniciativa.

A universidade também disponibilizará avançados recursos tecnológicos para o desenvolvimento do projeto, como acesso a servidores computacionais e softwares necessários. Além disso, a biblioteca da universidade oferece acesso a um vasto número de periódicos científicos.

Em relação à infraestrutura, a University of Exeter oferecerá um espaço dedicado no campus para o Dr. Chico Camargo, bem como amplo acesso a salas de reunião para colaborações e

apresentações. Esses espaços físicos são fundamentais para promover o trabalho colaborativo e a disseminação dos resultados da pesquisa. Para assegurar o bom andamento do projeto, a universidade proverá serviços de suporte abrangentes, incluindo assistência administrativa para gestão de projetos e elaboração de relatórios, além de suporte de TI para manutenção de equipamentos e suporte técnico contínuo.

13. Considerações Finais

Este projeto aborda uma das principais ameaças ao ecossistema de mídia contemporâneo: a manipulação da opinião pública através de comportamentos coordenados nas plataformas digitais. No Brasil, o YouTube é acessado diariamente por dezenas de milhões de pessoas, que utilizam a plataforma para busca de informação e orientações comportamentais. Esses usuários ficam vulneráveis a iniciativas que utilizam o CIB para manipular a percepção social sobre o debate público, amplificar desinformação e influenciar algoritmos de recomendação, interferindo no que as pessoas assistem e acreditam.

Além disso, o projeto traz avanços na aplicação de inteligência artificial para detectar padrões de manipulação, ajudando a identificar campanhas coordenadas que escapam dos métodos de moderação tradicionais. Com isso, ele oferece uma base científica para o desenvolvimento de métodos de identificação de estratégias de desinformação no Brasil.

Finalmente, ao fornecer recomendações práticas para reguladores e plataformas, o projeto pode contribuir diretamente para o combate à desinformação, protegendo a integridade do discurso público e promovendo um consumo de informação mais autêntico e plural. Isso é especialmente crucial em tempos de polarização política e disseminação de “fake news”, que ameaçam a democracia e o bem-estar social, a fim de garantir um ambiente digital mais íntegro, transparente e seguro para os usuários brasileiros.

14. Referências Bibliográficas

Alberto, T. C., Lochter, J. V.; Almeida, T. A. (2015) Tubespam: Comment spam filtering on youtube. In: 2015 IEEE 14th international conference on machine learning and applications (ICMLA). IEEE, p. 138-143.

Barros, C. E., Silva, D., Santos, M. L., Medeiros, P. M., Salles, D. G., Santini, R. M. (2024). Negacionismo Climático no Youtube: como argumentos de falsos especialistas repercutem nos comentários da audiência. 33º Congresso da Associação Nacional dos Programas de Pós-Graduação em Comunicação (ComPós), julho 2024.

<https://proceedings.science/compos/compos-2024/trabalhos/negacionismo-climatico-no-youtube-come-argumentos-de-falsos-especialistas-reperc#>

Bernardi, A.J., Bonami, B., Camargo, C., Bernardino, I. I., Richer, R. & Tsuzuki, C. (2022). Viu Política: Mapeando o conteúdo político do YouTube nas eleições brasileiras de 2022. *Instituto Vero*. <https://www.vero.org.br/viupolitica>

Bruns, A. (2021). Echo chambers? Filter bubbles? The misleading metaphors that obscure the real problem. In M. Pérez-Escobar & J. M. Noguera-Vivo (Eds.), *Hate speech and polarization in participatory society* (pp. 33-48). Routledge. <https://doi.org/10.4324/9781003109891-4>

Bossetta, M. (2020). Scandalous design: How social media platforms' responses to scandal impacts campaigns and elections. *Social Media + Society*, 6(2), 2056305120924777.

<https://journals.sagepub.com/doi/epub/10.1177/2056305120924777>

Cialdini, R. B. (2013). *Influence: Pearson new international edition* (5th ed.). London, England: Pearson Education.

Ciodaro, T., Carmo, V.D., Ferreira, F., Grael, F., Salles, D., & Santini, M. (2024). Visual Social Network Analysis Based on Deep-Walk Graph-Embeddings and Self-Organizing Maps. *Anais do XIII Brazilian Workshop on Social Network Analysis and Mining (BraSNAM 2024)*.

<https://sol.sbc.org.br/index.php/brasnam/article/view/29349/29154>

de-Lima-Santos, M. F., & Ceron, W. (2023a). Coordinated Amplification, Coordinated Inauthentic Behaviour, Orchestrated Campaigns: A Systematic Literature Review of

Coordinated Inauthentic Content on Online Social Networks. *Mapping Lies in the Global Media Sphere*, 165–184. <https://doi.org/10.4324/9781003403203-14>

de-Lima-Santos, M. F., & Ceron, W. (2023b). Disinformation echo chambers on Facebook. In *Fighting fake facts* (pp. 61–90). MDPI Books. <https://doi.org/10.3390/books978-3-0365-1347-8-4>

Ceron, W., de-Lima-Santos, M. F., & Quiles, M. G. (2021a). Fake news agenda in the era of COVID-19: Identifying trends through fact-checking content. *Online social networks and media*, 21. <https://doi.org/10.1016/j.osnem.2020.100116>

Ceron, W., Gruszynski Sanseverino, G., & de-Lima-Santos, M. F. (2021b). COVID-19 fake news diffusion across Latin America. *Social Network Analysis and Mining*, 11(47). <https://doi.org/10.1007/s13278-021-00753-z>

Felipe, M. (2022) Facebook aplica em usuários humanos cartilha usada contra robôs. *desinformante. <https://desinformante.com.br/facebook-aplica-em-usuarios-humanos-cartilha-usada-contra-robos/>

Hovy, D., & Prabhumoye, S. (2021). Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(8). <https://doi.org/10.1111/lnc3.12432>

Gallwitz, F., & Kreil, M. (2021). The Rise and Fall of “Social Bot” Research (SSRN Scholarly Paper 3814191). <https://papers.ssrn.com/abstract=3814191>

Giglietto, F., Righetti, N., Rossi, L., & Marino, G. (2020a). It takes a village to manipulate the media: Coordinated link sharing behavior during 2018 and 2019 Italian elections. *Information, Communication & Society*, 23(6), 867–891. <https://doi.org/10.1080/1369118X.2020.1739732>

Giglietto, F., Righetti, N., Rossi, L., & Marino, G. (2020b). Coordinated link sharing behavior as a signal to surface sources of problematic information on Facebook. *ACM International Conference Proceeding Series*, 20, 85–91. <https://doi.org/10.1145/3400806.3400817>

Giglietto, F., Iannelli, L., Rossi, L., Valeriani, A., Righetti, N., Carabini, F., Marino, G., Usai, S., & Zurovac, E. (2018). Mapping Italian news media political coverage in the lead-up to the 2018 general election. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.3179930>

González-Bailón, S., & Paltoglou, G. (2015). Signals of Public Opinion in Online Communication: A Comparison of Methods and Data Sources. *The ANNALS of the American Academy of Political and Social Science*, 659(1), 95–107.

<https://doi.org/10.1177/0002716215569192>

Hale, S. A., Belisario, A., Mostafa, A. N., & Camargo, C. (2024). Analyzing misinformation claims during the 2022 Brazilian general election on WhatsApp, Twitter, and Kwai. *International Journal of Public Opinion Research*, 36(3).

<https://academic.oup.com/ijpor/article/36/3/edae032/7709027>

Kantar IBOPE Media. (2024). Inside Video 2024. Abril de 2024.

<https://kantaribopemedia.com/inside-video-2024/>

Mannocci, L., Mazza M., Monreale, A., Tesconi, M., Cresci, S. (2024) Detection and Characterization of Coordinated Online Behavior: A Survey. arXiv preprint arXiv:2408.01257.

<https://arxiv.org/pdf/2408.01257>

McCrosky, J. & Geurkink, B. (2021). YouTube Regrets: A crowdsourced investigation into YouTube's recommendation algorithm. Mozilla Foundation.

<https://foundation.mozilla.org/en/youtube/findings/>

Möller, J. (2021). Filter bubbles and digital echo chambers. In H. Tumber, & S. Waisbord (Eds.), *The routledge companion to media disinformation and populism* (pp. 92-100). (Routledge Media and Cultural Studies Companions). Routledge.

<https://doi.org/10.4324/9781003004431-10>

Na, Seung & Cho, Sumin & Seungwon, Shin. (2023). Evolving Bots: The New Generation of Comment Bots and their Underlying Scam Campaigns in YouTube. 297–312.

<https://doi.org/10.1145/3618257.3624822>

NetLab UFRJ. [s.d]. NetLab na Imprensa. <https://netlab.eco.ufrj.br/imprensa>

NetLab UFRJ. (2024). Ecossistema de Desinformação Socioambiental no Brasil. Grupo de Trabalho de Economia Digital, G20 Brasil 2024, Painel Integridade da informação como fundamento da luta contra a mudança climática, 1 de maio de 2024.

<https://netlab.eco.ufrj.br/post/ecossistema-de-desinforma%C3%A7%C3%A3o-socioambiental-no-brasil>

Ntoutsis, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdil, W., Vidal, M.-E., Ruggieri, S., Turini, F., Papadopoulos, S., Krasanakis, E., Kompatsiaris, I., Kinder-Kurlanda, K., Wagner, C., Karimi, F., Fernandez, M., Alani, H., Berendt, B., Kruegel, T., Heinze, C., ... Staab, S. (2020). Bias in data-driven artificial intelligence systems—An introductory survey. *WIRES Data Mining and Knowledge Discovery*, 10(3), e1356. <https://doi.org/10.1002/widm.1356>

Pires, R., Abonizio, H., Almeida, T.S., Nogueira, R. (2023). Sabiá: Portuguese Large Language Models. In: Naldi, M.C., Bianchi, R.A.C. (eds) *Intelligent Systems. BRACIS 2023. Lecture Notes in Computer Science()*, vol 14197. Springer, Cham.

https://doi.org/10.1007/978-3-031-45392-2_15

Rauchfleisch, A., & Kaiser, J. (2020). The False positive problem of automatic bot detection in social science research. *PLOS ONE*, 15(10), e0241045.

<https://doi.org/10.1371/journal.pone.0241045>.

Rocha, S. S.; Silva, C. H. V.; Ferreira, C. H. G.; Gonçalves, G. D.; Almeida, J. M. Monitorando a Opinião Pública sobre Operações Policiais no Brasil via Comentários de Vídeos no YouTube. In: *Brazilian Workshop On Social Network Analysis And Mining (BRASNAM)*, 13. , 2024, Brasília/DF. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2024. p. 158-171. ISSN 2595-6094. DOI: <https://doi.org/10.5753/brasnam.2024.3101>

Salles, D., de Medeiros, P. M., Santini, R. M., & Barros, C. E. (2023). The Far-Right Smokescreen: Environmental Conspiracy and Culture Wars on Brazilian YouTube. *Social Media + Society*, 9(3). <https://doi.org/10.1177/20563051231196876>

Salles, D., Muniz de Medeiros, P., Martins, B., Regattieri, L., & Santini, R. M. (2024). The role of social bots in the Brazilian environmental debate: an analysis of the 2020 Amazon Forest fires in Twitter. *The International Review of Information Ethics*, 33(1).

<https://doi.org/10.29173/irie510>

Santini, R. M., Salles, D., Tucci, G., Ferreira, F., & Graef, F. (2020). Making up Audience: Media Bots and the Falsification of the Public Sphere. *Communication Studies*, 71(3), 466–487.

<https://doi.org/10.1080/10510974.2020.1735466>

Santini, R. M., Salles, D., Tucci, G. (2021) When Machine Behavior Targets Future Voters: The Use of Social Bots to Test Narratives for Political Campaigns in Brazil. *International Journal of Communication*, [S.l.], v. 15, p. 24, feb. 2021. ISSN 1932-8036.

<https://ijoc.org/index.php/ijoc/article/view/14803>

Santini, R. M., Salles, D., & Mattos, B. (2023). Recommending instead of taking down: Youtube hyperpartisan content promotion amid the Brazilian general elections. *Policy & Internet*, 15, 512–527. <https://doi.org/10.1002/poi3.380>

Santini, R. M., Salles, D., Ferreira, F., & Graef, F. (2023). Gotcha Bot Detection: Context, Time And Place Matters. In *SciELO Preprints*. <https://doi.org/10.1590/SciELOPreprints.5974>

Santini, R. M., Fitzgerald, J., Filho, H. R., Lokmanoglu, A. D., Sombatpoonsiri, J., Salles, D. (2024). The Economic and Social Impact of Platform Transparency: Ad Regulation for the Construction of Digital Market. Task Force 05 - Inclusive Digital Transformation, G20 Brasil 2024.

https://t20brasil.org/media/documentos/arquivos/TF05_ST_06_The_Economic_and_So66cf6bee8d2ed.pdf

Sen, I., Flöck, F., Weller, K., Weiß, B., & Wagner, C. (2021). A Total Error Framework for Digital Traces of Human Behavior on Online Platforms. *Public Opinion Quarterly*, 85(S1), 399–422.

<https://doi.org/10.1093/poq/nfab018>

Schuchard, R., Crooks, A., Stefanidis, A., & Croitoru, A. (2019a). Bots fired: Examining social bot evidence in online mass shooting conversations. *Palgrave Communications*, 5(1), 1–12.

<https://doi.org/10.1057/s41599-019-0359-x>

Schuchard, R., Crooks, A. T., Stefanidis, A., & Croitoru, A. (2019b). Bot stamina: Examining the influence and staying power of bots in online social networks. *Applied Network Science*, 4(1), Article 1.

<https://doi.org/10.1007/s41109-019-0164-x>

Souza, F., Nogueira, R., Lotufo, R. (2020). BERTimbau: Pretrained BERT Models for Brazilian Portuguese. In: Cerri, R., Prati, R.C. (eds) *Intelligent Systems. BRACIS 2020. Lecture Notes in Computer Science()*, vol 12319. Springer, Cham.

https://doi.org/10.1007/978-3-030-61377-8_28

TeBlunthuis, N., Hase, V., & Chan, C.-H. (2024). Misclassification in Automated Content Analysis Causes Bias in Regression. Can We Fix It? Yes We Can! *Communication Methods and Measures*, 0(0), 1–22. <https://doi.org/10.1080/19312458.2023.2293713>

Terren, L., & Borge, R. (2021). Echo chambers on social media: A systematic review of the literature. *Review of Communication Research*, 9, 1–39.

Wardle, C. & Derakhshan, H. (2017). *Information Disorder: Toward an interdisciplinary framework for research and policy making*, Council of Europe Report. <https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-research/168076277c>



NETLAB.ECO.UFRJ.BR