Web Scraping em Redes Sociais Digitais: um experimento com Selenium e BeautifulSoup frente ao fechamento das APIs[1]

Raquel Lobão Evangelista[2]
Universidade do Estado do Rio de Janeiro – UERJ
Cristiano do Nascimento[3]
Universidade Federal Fluminense -UFF

RESUMO

O avanço da dataficação impôs desafios à pesquisa em Comunicação, especialmente nos estudos de recepção em ambientes digitais. Com o fechamento das APIs das redes sociais na intenet, pesquisadores têm enfrentado obstáculos metodológicos para acessar dados essenciais à compreensão das práticas de interação e circulação de sentidos nas plataformas. Este trabalho discute o uso das ferramentas *Selenium* e *BeautifulSoup* como alternativas viáveis para a coleta de dados em redes sociais digitais. As ferramentas serão testadas no Instagram, no Facebook, no X (antigo Twitter) e no YouTube, com o objetivo de verificar em quais dessas plataformas elas funcionam adequadamente, possibilitando a construção de bancos de dados compostos por comentários, postagens e perfis. Fundamentado em autores como Van Dijck (2018), Bucher (2018) e Srnicek (2017), este experimento técnico aponta caminhos para uma abordagem eficiente, porém, crítica da pesquisa digital em tempos de plataformização.

PALAVRAS-CHAVE

dataficação; algoritmos; *web scrapping*; métodos digitais

CORPO DO TEXTO

A comunicação digital contemporânea é profundamente influenciada por processos de dataficação e plataformização, que convertem interações sociais em dados quantificáveis e reconfiguram as dinâmicas de produção e consumo de conteúdo. A dataficação é entendida por Van Dijck (2018) como a transformação de aspectos da vida cotidiana em dados digitais, permitindo seu monitoramento e análise. Para Couldry e Mejias (2019), trata-se de um projeto extrativista global que busca "transformar a vida humana em matéria-prima gratuita para processos de comercialização" (p. 3). Essa mudança é mediada por plataformas digitais que operam como modelos de negócio centrados na extração e monetização de dados.

---

[1] Trabalho apresentado no Grupo de Trabalho Metodologias interdisciplinares para análises em ambientes digitais, evento integrante da programação do 28º Congresso de Ciências da Comunicação na Região Sudeste, realizado de 15 a 17 de maio de 2025.

[2] Professora Adjunta no Programa de Pós-Graduação da UERJ e no Departamento de Jornalismo. E-mail: raquel.lobao@uerj.br

[3] Mestrando no Programa de Ciências da Computação. E-mail: cristiano_nascimento@id.uff.br

Seguindo este pensamento, Carlos d'Andréa (2020) destaca que as plataformas digitais possuem especificidades políticas e materiais que moldam suas operações, incluindo seus modelos de negócio, que são fundamentais para compreender as dinâmicas contemporâneas da comunicação digital. Nesse contexto, emerge o conceito de capitalismo de vigilância, conforme discutido por Shoshana Zuboff (2019), que descreve um novo modelo econômico baseado na coleta massiva de dados pessoais para prever e influenciar comportamentos futuros. Diferentemente do capitalismo tradicional, em que a exploração se dava sobre a força de trabalho, no capitalismo de vigilância, a exploração recai sobre a própria experiência humana, capturada e convertida em dados pelas plataformas digitais. Esse modelo levanta preocupações éticas e políticas significativas, pois transforma aspectos íntimos da vida em *commodities.*

Nesse contexto de plataformização e dataficação intensificadas, os pesquisadores da área da Comunicação, especialmente aqueles dedicados aos estudos de recepção, viram-se diante da necessidade de adaptar suas abordagens metodológicas. Com a centralidade crescente dos ambientes digitais na vida cotidiana, novos objetos de estudo emergiram - como interações em comentários, compartilhamentos, algoritmos de recomendação e práticas de engajamento em redes sociais - que não podiam mais ser plenamente compreendidos por meio de métodos clássicos como entrevistas, questionários ou grupos focais.

O volume e a complexidade dos dados gerados nesses ambientes exigiram o desenvolvimento de novas competências analíticas e técnicas. Bancos de dados robustos passaram a fazer parte da realidade investigativa, exigindo práticas mais automatizadas de coleta e tratamento de dados. Foi nesse cenário que muitos pesquisadores passaram a utilizar APIs (*Interfaces de Programação de Aplicações*), que são conjuntos de códigos e instruções disponibilizados pelas plataformas para permitir o acesso estruturado a dados públicos. Essas interfaces permitiram, por um tempo, extrair informações diretamente das redes sociais na internet de forma legítima e organizada, viabilizando análises em larga escala. No entanto, como aponta Zuboff (2019), as plataformas digitais reconheceram que o controle sobre esses dados representava um ativo econômico estratégico, e, por isso, começaram a restringir ou encerrar o acesso a essas APIs públicas. Esse movimento aumentou ainda mais a assimetria entre empresas e pesquisadores, tornando o acesso a dados uma questão de poder e rentabilidade: “o excedente comportamental se tornou a

principal fonte de receita das grandes plataformas" (ZUBOFF, 2019, p. 98). Ao restringir o acesso de pesquisadores, jornalistas e desenvolvedores independentes, essas empresas reforçam sua posição monopolista sobre os fluxos informacionais e ampliam o valor comercial dos dados que capturam. Essa prática está diretamente ligada ao funcionamento do capitalismo de vigilância, como descrito por Zuboff (2019), pois "quem detém os dados detém o poder de prever e modificar o comportamento humano em escala" (p. 8). Trata-se de um movimento de cercamento digital, em que o acesso ao conhecimento sobre as interações sociais é bloqueado em nome da extração exclusiva de valor pelas plataformas.

Como consequência, hoje, os pesquisadores enfrentam um cenário de escassez informacional, dependência tecnológica e opacidade estrutural, no qual a coleta de dados tornou-se mais difícil, cara e politicamente sensível. Diante desse cenário, impõe-se a seguinte pergunta de pesquisa: como os pesquisadores em Comunicação, especialmente na área de recepção, podem contornar as limitações impostas pelo fechamento das APIs das plataformas digitais e continuar acessando dados relevantes para suas investigações? A partir dessa questão, este trabalho propõe as seguintes hipóteses: 1) o uso de métodos computacionais alternativos e técnicas de raspagem de dados pode oferecer caminhos viáveis para a coleta de informações em ambientes digitais, mesmo diante do fechamento das APIs; 2) a restrição de acesso a dados por parte das plataformas digtiais redefine os próprios objetos de estudo da recepção, exigindo novos olhares interdisciplinares; 3) a opacidade algorítmica e a centralização dos dados pelas plataformas digitais configuram não apenas um desafio técnico, mas uma disputa epistemológica que exige resistência metodológica e crítica política por parte da pesquisa em Comunicação. Portanto, aqui, o objetivo geral é analisar em que medida a adoção de ferramentas computacionais de raspagem de dados podem representar uma resposta viável às restrições impostas pelas plataformas digitais à coleta de dados, contribuindo para a adaptação metodológica da pesquisa em recepção.

A raspagem de dados ou *web scraping* é um conjunto de técnicas voltadas para a extração automatizada de dados de páginas web. Elas permitem que estruturas de informação visíveis ao usuário - como textos, imagens, links e comentários - sejam coletadas por meio de *scripts* que simulam a navegação e interpretam o conteúdo HTML das páginas. Tais técnicas se tornaram especialmente relevantes após o fechamento

progressivo das APIs públicas pelas plataformas digitais, tornando o *scraping* uma alternativa metodológica para pesquisadores interessados em acessar dados disponibilizados publicamente na interface das redes sociais.

Sua aplicação requer o uso de ferramentas e bibliotecas de dados e códigos específicas que variam em complexidade e finalidade. Algumas soluções operam por interface gráfica, permitindo que usuários sem conhecimentos de programação realizem coletas simples; outras, mais robustas, dependem de código personalizado e oferecem maior controle sobre o processo. Entre os desafios técnicos mais comuns está a manipulação de páginas que utilizam *JavaScript* para carregar conteúdo dinamicamente - situação que exige ferramentas capazes de renderizar o conteúdo como navegadores reais. Nesse sentido, a escolha da ferramenta adequada deve considerar o tipo de dado desejado, a estrutura da página e a frequência de coleta.

Considerando estes aspectos, procuramos inserir este experimento em uma abordagem qualitativa, de cunho exploratória-descritiva, com foco na coleta e análise de dados disponíveis a partir das ferramentas programáticas *Selenium* e *BeautifulSoup.* Elas foram testadas para minerar postagens, comentários e perfis no Instagram, Facebbok, X e Youtube.

Ainda que brevemente, vale mencionar que o *BeautifulSoup* é uma biblioteca em *Python* destinada à extração de dados de documentos HTML e XML. Sua principal função é permitir a leitura e estruturação dos elementos de uma página web de maneira simples, oferecendo recursos para localizar, navegar e modificar árvores de elementos HTML. Já o *Selenium* é uma ferramenta de automação de navegadores, que simula o comportamento humano em ambientes web. Com ela, é possível interagir com páginas que requerem carregamento dinâmico de conteúdo via *JavaScript*, algo comum no Instagram e em outras redes sociais contemporâneas.

Do ponto de vista ético, o *web scraping* levanta uma série de questões que precisam ser cuidadosamente avaliadas pelos pesquisadores. Embora o acesso a dados públicos seja legal em muitos contextos, a extração em massa, a sobrecarga de servidores e o uso de dados sensíveis sem consentimento podem configurar violações de boas práticas e diretrizes institucionais de pesquisa. Como reforçado no material analisado, é essencial realizar coletas respeitando os limites estabelecidos pelas próprias plataformas, observar as condições de uso dos sites e adotar posturas responsáveis quanto à

anonimização e à proteção dos dados coletados. Nesse sentido, o uso do *web scraping* em pesquisa acadêmica exige domínio técnico e consciência crítica.

REFERÊNCIAS

BUCHER, Taina. **If...Then: Algorithmic Power and Politics**. Oxford: Oxford University Press, 2018.

COULDRY, Nick; MEJIAS, Ulises A. **The Costs of Connection: How Data Is Colonizing Human Life and Appropriating It for Capitalism**. Stanford: Stanford University Press, 2019.

D'ANDRÉA, Carlos. Plataformas digitais e disputas sociotécnicas: notas para uma agenda de pesquisa. *Revista Brasileira de Ciências da Comunicação*, v. 43, n. 2, 2020. Disponível em: https://www.scielo.br/j/interc/a/7pWBPfVwRZnLMLHQcsFQ7Bp/?lang=pt. Acesso em: 31 mar. 2025.

SRNICEK, Nick. **Capitalismo de plataforma**. São Paulo: Autonomia Literária, 2017.

VAN DIJCK, José. **A cultura da conectividade: Uma história crítica das redes sociais**. São Paulo: Paulus, 2018.

ZUBOFF, Shoshana. **A era do capitalismo de vigilância: A luta por um futuro humano na nova fronteira do poder**. Rio de Janeiro: Intrínseca, 2019