Um pipeline multiplataforma unificado para coleta e padronização de dados de mídias sociais

Anderson Frasão
aacfrasao@inf.ufpr.br
Universidade Federal do Paraná
Curitiba, Paraná, Brasil

Tiago Heinrich
theinric@mpi-inf.mpg.de
Max Planck Institute for Informatics
Saarbrücken, Saarland, Germany

Vinicius Fulber-Garcia
vinicius@inf.ufpr.br
Universidade Federal do Paraná
Curitiba, Paraná, Brasil

Abstract

Social networks have become a central source of data for large-scale empirical research. However, access to this data is constrained by official API limitations, opaque commercial solutions, and fragmented scraping tools. This work presents a tool for collecting and storing multi-platform data on social networks, designed as an integrated pipeline that includes Instagram, TikTok, Twitter/X, and YouTube. The tool combines automated collection with structured extraction of text, images, videos, audio, metadata, URLs, and comments, organizing these elements into a standardized relational schema with tables for users, media, and comments on each platform. This standardization reduces the engineering effort required to consolidate heterogeneous data, facilitating comparative analyses between networks and enabling the application of multimodal analysis and Natural Language Processing (NLP) methods in different research scenarios, including studies of fraud, engagement, and abusive behavior. In addition, the pipeline incorporates time-based controls, logging, and multi-account and cookie management, making the collection process more robust in the face of blocks, access limits, and platform changes. The tool thus aims to serve as reusable infrastructure for sociotechnical research on social media, promoting reproducibility and transparency during data collection.

Keywords

REDES SOCIAIS, COLETA DE DADOS, *WEB SCRAPING*.

1 Introdução

As redes sociais se converteram, nas últimas duas décadas, em um dos principais ambientes de interação humana e produção contínua de dados digitais. Esse crescimento acompanha o avanço do fenômeno de *big data*, no qual grandes volumes de informações são gerados em alta velocidade e variedade, permitindo observar fenômenos sociais em escala inédita. Conforme discutido em Schroeder [1], essas plataformas tornaram-se fonte estratégica para pesquisas sobre comportamento coletivo, influências sociais e processos comunicacionais.

Além do volume, redes sociais se destacam pela diversidade multimodal de dados, incluindo textos, imagens, vídeos e metadados. Ghani et al. [2] destaca como essa variedade possibilita estudos sobre difusão de conteúdo, engajamento, predição de eventos e análise de sentimentos, ao integrar técnicas computacionais, estatísticas e linguísticas. Além disso, a coleta desses dados pode alavancar pesquisas relacionadas à detecção, prevenção e mitigação dos diferentes tipos de golpes e fraudes que ocorrem no contexto dessas redes [3]. Em paralelo, esforços de pesquisa passaram a adotar metodologias mais robustas para capturar e analisar esses dados, impulsionando estudos sobre redes complexas e comunicação digital.

Considerando o cenário descrito, cresce a busca e o uso de ferramentas de coleta automatizada de dados em redes sociais para as mais diferentes finalidades. Com isso, APIs, bibliotecas de extração e rotinas de automação tornaram-se essenciais para suprir a demanda por dados em larga escala e lidar com as características e limitações de cada plataforma em particular [4]. Ou seja, um ambiente digital altamente dinâmico exige mecanismos capazes de coletar, estruturar e armazenar informações provenientes de múltiplas fontes, com diferentes formatos.

Porém, a redução no número e as limitações impostas pelas ferramentas oficiais para coleta de dados em redes sociais se tornaram pontos centrais de debate na comunidade acadêmica recentemente. No contexto descrito como era "pós-API" e "pós-pós-API", o fechamento de APIs públicas e o fim de programas acadêmicos gratuitos, como no Twitter/X e no CrowdTangle, reduziram drasticamente os canais formais de acesso a dados [5]. Mesmo apoiando-se em iniciativas regulatórias, como o Digital Services Act, os pesquisadores enfrentam processos de credenciamento complexos e lidam com APIs pouco funcionais para acessar os dados desejados [6]. Isso reforça uma condição de "independência sob permissão", na qual a viabilidade das pesquisas depende das interpretações das plataformas sobre o que pode ser acessado [6].

Além das barreiras de acesso, também existem limitações técnicas e epistemológicas nas APIs. Auditorias recentes da Research API do TikTok mostram que parte relevante dos metadados de conteúdos publicamente acessíveis não é retornada: cerca de um em cada oito vídeos de um conjunto doado por usuários não pôde ser recuperado, sem justificativa clara [7]. Isso compromete critérios clássicos de qualidade de dados e coloca em risco a confiabilidade de estudos que dependem exclusivamente desses canais. Problemas semelhantes ocorrem em outras plataformas, seja por *rate limiting*, amostragem não documentada ou filtragens invisíveis, dificultando saber o que o conjunto de dados representa. Autores como Brooker et al. [8] argumentam que essas restrições moldam as análises possíveis e devem ser consideradas na interpretação dos resultados.

Diante dessas limitações, os pesquisadores recorrentemente executam processos de coleta manual ou semi-automatizada, como *scraping* com Selenium ou ferramentas próprias. Porém, o volume, a velocidade e a dinamicidade dos dados tornam esses métodos difíceis de escalar e reproduzir, especialmente sem ferramentas consolidadas ou padrões amplos [9, 10]. Pesquisas etnográficas com cientistas de dados mostram que muitas soluções de contorno, como ajustes improvisados em *scripts* ou uso de recursos não canônicos para contornar limites de API, são raramente documentadas em detalhes, o que prejudica a replicabilidade [11].

Nesse contexto de acesso escasso, APIs incompletas e coleta manual custosa, cresce a relevância do desenvolvimento de técnicas e

ferramentas que automatizem, integrem e documentem de forma transparente a coleta multiplataforma. É nesse cenário que se insere a proposta apresentada neste trabalho, objetivando o estabelecimento de uma ferramenta voltada à coleta integrada, multimodal e multiplataforma de dados em redes sociais. Essa ferramenta deriva do desenvolvimento e implementação de um *pipeline* unificado para a coleta de dados, aplicado, neste trabalho, às plataformas Instagram, Twitter/X, YouTube e TikTok. Tal *pipeline* integra rotinas que extraem informações essenciais sobre perfis, publicações e interações, incluindo comentários, relações de engajamento e outros elementos relevantes para a comunidade acadêmica que atua na área.

Além da coleta multiplataforma, uma contribuição central está na extração estruturada de diferentes tipos de dados, como textos, vídeos, imagens, áudios, metadados, URLs e comentários, organizados em um banco padronizado projetado para facilitar análises posteriores. A arquitetura da ferramenta foi concebida para ser reprodutível e expansível, permitindo a incorporação de novos módulos para outras plataformas de redes sociais ou tipos de dados. Assim, o artigo apresenta uma infraestrutura metodológica e técnica que favorece transparência, escala e redução do esforço técnico na coleta de dados, permitindo que os pesquisadores concentrem energia nas etapas analíticas, em vez do processo de coleta.

O restante deste artigo está organizado da seguinte forma: a Seção 2 apresenta os trabalhos relacionados. A Seção 3 apresenta a metodologia utilizada nesse trabalho. A Seção 4 apresenta a implementação da nossa ferramenta. A Seção 5 apresenta estudos de caso para nossa ferramenta. A Seção 6 apresenta as considerações éticas em relação à coleta de dados. Por fim, a Seção 7 traz as considerações finais e expectativas futuras em relação ao trabalho.

2 Trabalhos Relacionados

As soluções existentes para coleta de dados em mídias sociais incluem serviços comerciais, APIs oficiais e ferramentas de *scraping* abertas, mas todas apresentam limitações claras. Plataformas como Make, Bright Data, Apify e Octoparse oferecem automação e *scrapers* prontos para o uso, embora funcionem como "caixas pretas", com pouca transparência e dependência de modelos pagos [12–16]. Já as APIs oficiais de TikTok e Twitter/X operam com credenciamento restritivo, limites de requisição e planos pagos, o que reduz a acessibilidade [17, 18]. Ferramentas abertas como Instaloader, snscrape e yt-dlp ampliam o alcance da coleta, mas carecem de padronização, pois cada uma produz saídas próprias e sem integração nativa entre plataformas [19–21]. Trabalhos acadêmicos reforçam que a coleta costuma ser fragmentada, dependente de *pipelines* específicos por rede e sujeita a instabilidades, mudanças silenciosas e heterogeneidade de formatos, o que aumenta o esforço de consolidação e dificulta análises comparativas [22–30].

Essas pesquisas apresentam duas conclusões particularmente convergentes: a coleta distribuída por ferramentas independentes exige integração manual, e a ausência de um *pipeline* unificado limita análises mais amplas. Estudos recentes exploram diferentes enfoques analíticos, como a estruturação de dados do Instagram para investigações de interesse público, avaliação de campanhas de *marketing*, estudo de comportamento digital e análise de engajamento durante a COVID-19. No TikTok, pesquisas combinam *scraping*, API oficial e modelos teóricos de dados para examinar redes sociais, interações e dinâmicas relacionais. Para o Twitter/X, o foco recai no *scraping* completo de *timelines*, no uso de múltiplos IPs e na coleta para alimentar classificadores de sentimento. No YouTube, análises exploram desempenho computacional da coleta, música, cultura e metadados de vídeos. Em conjunto, os trabalhos da literatura evidenciam que os tipos de análises já realizados vão de investigações sociais e culturais a modelagens de comportamento e estudos de redes, reforçando a necessidade de estruturas de coleta mais integradas, transparentes e expansíveis.

O *pipeline* proposto neste trabalho busca precisamente mitigar essas limitações ao oferecer uma arquitetura unificada de coleta multiplataforma, com módulos específicos para Instagram, TikTok, Twitter/X e YouTube conectados a uma camada comum de armazenamento. Os dados coletados são organizados em um banco de dados padronizado, com esquemas compatíveis entre redes para usuários, mídias e comentários, facilitando consultas comparativas e análises conjuntas. Além disso, o *pipeline* foi concebido para lidar com multimodalidade de forma estruturada, associando textos, arquivos de mídia e metadados em uma mesma unidade analítica, e para ser reprodutível e expansível, permitindo a adaptação a mudanças nas APIs e a inclusão de novas plataformas por meio de alterações pontuais e isoladas em módulos específicos da arquitetura/ferramenta, sem a necessidade de processos completos de reengenharia.

3 Metodologia

Esta seção descreve a arquitetura, o funcionamento e os componentes centrais da ferramenta de coleta desenvolvida. O objetivo é detalhar como os dados são obtidos das diferentes plataformas, como são extraídos os elementos multimodais, e como essas informações são organizadas de forma estruturada para análises subsequentes. A metodologia segue o *pipeline* apresentado na Figura 1, que organiza o processo em quatro blocos principais: (i) busca nas plataformas, (ii) extração de elementos multimodais, (iii) armazenamento em banco de dados e (iv) disponibilização dos dados para análises posteriores.

3.1 Visão Geral

A ferramenta foi concebida como um *pipeline* de coleta multiplataformas, ilustrado na Figura 1. Em um primeiro estágio, o pesquisador define o conjunto de perfis, canais ou publicações-alvo em cada rede social, a partir de um critério alinhado à pesquisa (por exemplo, contas institucionais, influenciadores, marcas específicas ou *hashtags*). Em seguida, módulos de coleta independentes são acionados para o Instagram, Twitter/X, YouTube e TikTok, responsáveis por interagir com cada plataforma, recuperar os conteúdos de interesse e registrar os respectivos metadados.

Os dados brutos retornados pela plataforma são então convertidos em um conjunto de entidades comuns (usuários, mídias e comentários), de modo a reduzir as diferenças de formato entre as redes. Essa informação é armazenada em um banco de dados relacional, que atua como camada central de consolidação. A partir dessa base padronizada, é possível derivar diferentes recortes analíticos, exportar subconjuntos de dados para outras ferramentas (por exemplo, notebooks, *pipelines* de aprendizado de máquina ou

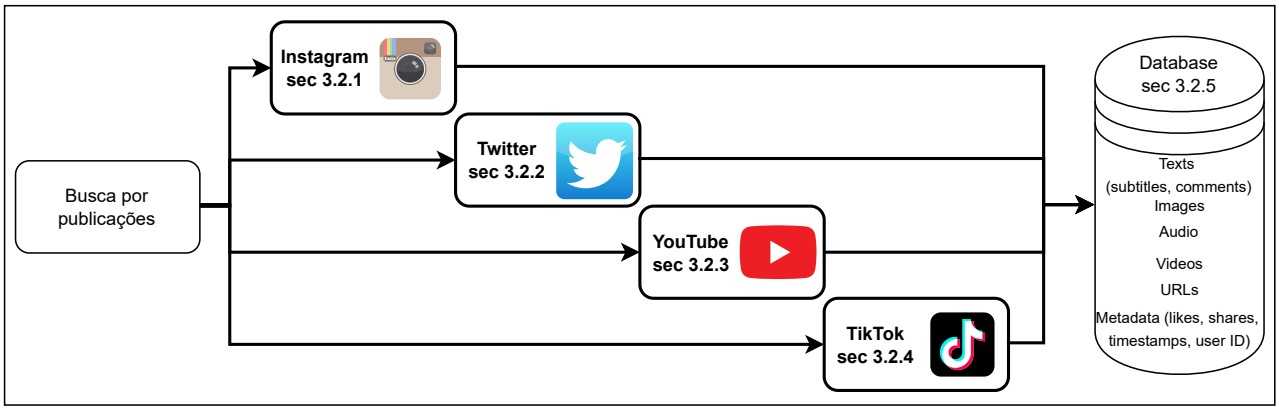

Figura 1: *Pipeline* Simplificado de Coleta de Dados

softwares de análise qualitativa) e reproduzir experimentos em novos contextos somente substituindo a lista de perfis ou o intervalo temporal de coleta.

O *pipeline* foi projetado para ser modular e expansível: cada módulo de coleta pode ser atualizado ou substituído sem comprometer os demais componentes. Essa separação permite que alterações nas APIs das plataformas, mudanças de *layout* ou ajustes de mecanismos anti-bot sejam tratados localmente em cada módulo.

3.2 Requisitos

A construção de um pipeline de coleta multiplataforma para mídias sociais pressupõe que se tenha acesso contínuo, estruturado e reprodutível aos dados fornecidos pela plataforma-alvo. Em consonância com trabalhos prévios que enfatizam os desafios de extração de informações em ambientes heterogêneos e com limitações de acesso [5, 9–11], este trabalho estabelece a necessidade de uma ferramenta capaz de integrar. padronizar e armazenar dados multimodais provenientes de diferentes redes sociais, de modo a superar restrições impostas por APIs. instabilidade de *scraping* e formatos divergentes de retorno.

Nesse sentido, a ferramenta deve atuar como um intermediário entre as plataformas e o pesquisador, abstraindo particularidades técnicas e fornecendo um conjunto de entidades padronizadas (usuário, mídia e comentários) independentemente da origem dos dados. Além disso, deve lidar com mecanismos de autenticação, cookies, múltiplas contas, limites de acesso e mudanças frequentes nas interfaces das plataformas, garantindo que a coleta seja consistente ao longo do tempo. Com base nessas necessidades, foram definidos os seguintes requisitos funcionais da ferramenta:

- **RF1** - Coletar dados multimodais (texto, imagem, vídeo, áudio, URLs e metadados) de cada plataforma suportada;
- **RF2** - Normalizar dados heterogêneos em um conjunto padronizado de entidades (usuários, medias e comentários);
- **RF3** - Registrar, para cada item coletado, os metadados essenciais (engajamento, datas, identificadores e vínculos entre entidades);
- **RF4** - Armazenar todos os dados em bancos relacionais organizados por plataforma, seguindo um *schema* lógico uniforme;
- **RF5** - Tolerar limitações de acesso impostas por APIs, cookies, bloqueios temporários e mecanismos anti-bot.

Por sua vez, os requisitos não funcionais priorizam robustez, modularidade, desempenho e reprodutibilidade. São eles:

- **RNF1** - Permitir a adição, remoção e atualização de módulos de coleta sem impacto nos demais componentes;
- **RNF2** - Manter uma interface interna unificada para operações de banco de dados em todas as plataformas;
- **RNF3** - Suportar a execução prolongada de coletas, com mecanismos de temporização e pausas configuráveis para evitar bloqueios;
- **RNF4** - Registrar todas as ações da ferramenta em logs padronizados, facilitando auditoria e depuração;
- **RNF5** - Garantir persistência confiável dos dados, evitando perdas por falhas de conexão, *timeout* ou interrupções inesperadas;
- **RNF6** - Ser compatível com ambiente Unix.

3.3 Módulos de Coleta

Cada rede social é tratada por um módulo específico, que implementa a lógica de acesso apropriada (API oficial, bibliotecas de terceiros ou *scraping* com navegador automatizado), mas todos seguem um mesmo modelo de processo: para cada publicação, o módulo registra informações sobre o autor, a mídia principal e os comentários associados, além de metadados numéricos e temporais. Esta subseção descreve, de forma resumida, quais informações são extraídas em cada plataforma.

3.3.1 Instagram. Para o Instagram, a coleta é realizada por meio da biblioteca `instagrapi`, que permite o acesso autenticado a perfis públicos ou semipúblicos. As mídias coletadas incluem fotos e vídeos publicados no *feed*. Para cada usuário, são armazenados *username*, nome completo, biografia, quantidade de mídias publicadas, número de seguidores, número de contas seguidas e *links*, presentes na biografia. Para cada mídia, são registrados a legenda textual, a data de publicação e o número de curtidas. Os comentários associados a cada publicação também são coletados, incluindo texto, data, número de curtidas e um subconjunto do perfil do autor do comentário (*username*, nome completo, biografia, número de

mídias, seguidores, seguindo e *links* presentes na bio). Esses dados são posteriormente normalizados e vinculados por identificadores de usuário e de mídia.

3.3.2 Twitter/X. Para o Twitter, a coleta é baseada em *scraping* com `Selenium`, que automatiza a navegação na interface web da plataforma, e `yt-dlp` para download de mídias. O módulo captura publicações que aparecem na página do perfil ou em consultas específicas, registrando tanto o conteúdo textual do *tweet* quanto as mídias anexadas (imagens, vídeos e áudios, quando disponíveis). Em relação ao autor, são extraídos o *username*, nome completo, a biografia, os *links* presentes no perfil e a data de criação da conta. Para cada *tweet*, são armazenados a descrição (texto), a data de publicação, número de respostas, número de *reposts*, número de curtidas e número de marcadores, quando disponível. Comentários (respostas) são coletados como instâncias adicionais do texto, associados ao *tweet* original e ao usuário, dono do *tweet*, preservando pelo menos *username* e nome completo para análise posterior de interações.

3.3.3 YouTube. No YouTube, a coleta é realizada por meio da `googleapiclient` e da YouTube Data API, complementada pelo download de *thumbnails* das publicações. Para cada canal, o sistema armazena título, descrição, número de visualizações agregadas, número de inscritos e quantidade de vídeos publicados. Ao nível de vídeo, são coletados títulos, descrições, data de publicação, número de visualizações, número de curtidas e número de comentários. A imagem da capa (*thumbnail*) é baixada e vinculada ao registro da mídia. Os comentários de cada vídeo são coletados com seu texto, nome do autor, data de publicação, número de curtidas e, quando disponível, o número de respostas. Dessa forma, o módulo do YouTube produz uma estrutura que conecta canal, vídeos e comentários, preservando tanto o conteúdo textual quanto elementos visuais relevantes para análise multimodal.

3.3.4 TikTok. Para o TikTok, a estratégia combina o uso do `TikTokApi` com o `Selenium` e `yt-dlp`, de modo a contornar limitações de acesso e lidar com conteúdo carregado dinamicamente. O módulo coleta tanto vídeos quanto publicações em formato de fotos, quando disponíveis. Em relação ao autor, são extraídos *username*, nome completo, biografia, número de mídias publicadas, número de seguidores e número de contas seguidas. Para cada mídia, registra-se a descrição textual e os principais contadores de engajamento disponibilizados na interface (como curtidas e, quando disponíveis, compartilhamentos e comentários). Os comentários são coletados com seu texto, número de curtidas e o *username* do autor, permitindo reconstruir interações em torno de cada publicação. As mídias (vídeos, imagens e áudios) são armazenadas em disco e referenciadas no banco de dados por caminho de arquivo associado ao identificador da publicação.

3.4 Elementos Multimodais Extraídos

Independentemente da plataforma de origem, os dados coletados são organizados em um conjunto de categorias teóricas multimodais comuns. A categoria *Texts* abrange legendas, postagens, descrições de vídeo, comentários e outros campos textuais, formando a base para tarefas de análise de conteúdo, extração de temas e análise de sentimentos. A categoria *Images* inclui fotos do Instagram e do TikTok, imagens anexas a *tweets* e *thumbnails* de vídeos do YouTube. Quando disponível, a categoria *Videos* contempla arquivos de vídeo baixados e associados às publicações originais, permitindo análise visual ou extração posterior de quadros. A categoria *Audio* inclui músicas e áudios anexados a imagens presentes nas publicações.

Além disso, o *pipeline* registra explicitamente URLs presentes em biografias e um conjunto de metadados, composto por contadores de engajamento (curtidas, respostas, comentários, compartilhamento, visualizações), carimbos temporais, identificadores internos das plataformas e relacionamento entre entidades (autor-mídia-comentário). Todos esses elementos multimodais são finalmente agregados em um esquema relacional padronizado, no qual usuários, mídias e comentários são tabelas principais conectadas por chaves estrangeiras. Essa organização facilita tanto consultas simples (por exemplo, engajamento médio por publicação) quanto análises mais complexas que combinem texto, imagens, vídeos e metadados em um mesmo experimento.

4 SoMeCollect: Ferramenta de Coleta Multiplataforma de Mídias Sociais

Nesta seção, o *pipeline* de execução da ferramenta de coleta multiplataforma é descrito, desde a definição das entradas até a consolidação dos dados em bancos relacionais. A implementação é organizada em módulos específicos por plataforma (`<sm>.py`[1]), apoiado por módulos compartilhados de temporização (`timer.py`), registro de logs (`logger.py`) e acesso unificado ao banco de dados (`database.py`)[2].

O **ponto de entrada** do *pipeline* é a definição das contas e conteúdos alvos em cada rede social. Para cada módulo de coleta, há uma lista de usuários, associando cada perfil a um conjunto de postagens a serem coletadas. Em seguida, funções do tipo `process_all()` percorrem essas estruturas e chamam iterativamente funções `process_video()` ou equivalentes, disparando a coleta por plataforma.

Na etapa de **coleta**, cada módulo implementa a lógica de interação com a respectiva plataforma. O módulo do Instagram utiliza a biblioteca `instagrapi` para autenticar múltiplas contas, lidar com desafios de login (`LoginRequired`, `ChallengeRequired`, `SubmitPhoneNumberForm`) e recuperar informações de perfis, postagens e comentários. Além de metadados como legenda, data, contagem de curtidas e tipo de mídia (foto, vídeo, álbum, IGTV, clips), o código faz o download dos arquivos de mídia para diretórios específicos (`../media/ig/<id>`), controladamente por esperas aleatórias e pausas longas gerenciadas pela classe `TimeManager`. No caso do Twitter/X, a coleta é feita via *scraping* com `Selenium`, utilizando um navegador `Chromium` automatizado. O módulo acessa diretamente a URL do *tweet*, extrai o texto, o autor, as mídias associadas (imagens hospedadas em `twimg.com/media`), as métricas de engajamento (respostas, reposts, curtidas, marcadores), tenta baixar imagens via `requests` ou baixar vídeos/áudios com `yt-dlp`, armazenando tudo em diretórios do tipo `../media/x/<id>`. Os comentários são coletados rolando a página e identificando novos artigos com conteúdo textual e dados de usuários.

[1] Considere $< sm > \in ig, tt, x, yt$.
[2] https://github.com/Carmofrasao/SoMeCollect

Para o YouTube, o módulo correspondente integra-se à API por meio do `googleapiclient`, recuperando metadados de vídeos (título, descrição, canal, data de publicações, contagem de visualizações, curtidas e comentários), bem como informações do canal (nome, descrição, total de visualizações, número de inscritos e quantidade de vídeos). A imagem de capa de cada vídeo (*thumbnail*) é baixada a partir da URL pública (`img.youtube.com`) e salva como `thumbnail.jpg` em diretórios do tipo `../media/yt/<id>`. Comentários são obtidos em páginas sucessivas da API, incluindo autor, texto, data, curtidas e, quando presentes, respostas encadeadas. No TikTok, o módulo combina `TikTokApi` com `Selenium` e `yt-dlp`, usando *cookies* para contornar limitações de acesso. São coletados perfis de usuários (nome, *username*, biografia, contagem de seguidores, seguindo e vídeos), além de comentários (texto, autor e contagem de curtidas). Os arquivos de mídia (vídeos/imagens/áudio) são baixados para diretórios do tipo `../media/tt/<id>` por meio de chamadas ao `yt-dlp` e de extração de URLs de imagens diretamente do DOM.

A etapa de **normalização** é realizada nos próprios módulos de coleta, que convertem as respostas brutas em um conjunto comum de entidades: usuário, mídia e comentários. Embora cada API ou página retorne estruturas distintas, o código mapeia campos específicos (por exemplo, seguidores vs. inscritos, curtidas vs. visualizações, descrições vs. legendas) para um vocabulário interno padronizado. Assim, dados de perfil são sempre representados com campos como *username*, nome exibido, biografia e contagens agregadas; mídias são representadas com identificadores estáveis (como *code*), texto associado (legenda ou descrição), caminhos de mídia em disco e contadores de engajamento; enquanto comentários são armazenados com texto, autor e metadados opcionais (data, curtida, respostas).

A **persistência em banco de dados** é centralizada no módulo `database.py`, que define a classe `Database`. Essa classe abre uma conexão `SQLite` independente para cada rede social (`../db/<sm>.db`) e mantém cursores associados a cada identificador de plataforma ("`ig`", "`tt`", "`x`", "`yt`"). As operações de inserção, consulta e atualização são abstraídas pelos métodos `insert()`, `request()`, `media_request()` e `update()`, que seguem uma convenção de nomenclatura de tabelas do tipo `<sm>_user_info`, `<sm>_media_info` e `<sm>_comments`. Dessa forma, todos os módulos de coleta utilizam a mesma interface para gravar e recuperar dados, impondo uma padronização de esquema lógico entre as plataformas, mesmo que cada uma utilize um arquivo de banco distinto.

Ao longo de todo o *pipeline*, a ferramenta utiliza um sistema de ***logging* e controle de tempo** compartilhado. O módulo `logger.py` configura *loggers* dedicados, silencia bibliotecas verbosas e garante que mensagens de depuração, informações, aviso e erro sejam emitidas consistentemente em todos os módulos. Já o módulo `timer.py` fornece funções de espera aleatória de curta, média e longa duração, além de um gerenciador de pausas periódicas (`TimeManager`) que introduz pausas prolongadas após janelas de execução configuráveis. Essa combinação reduz o risco de bloqueio por parte das plataformas e simula um padrão de acesso mais próximo do comportamento humano, especialmente na coleta intensiva de comentários e mídias.

Por fim, na etapa de **exploração e uso**, os bancos de dados resultantes funcionam como ponto único de acesso aos dados coletados. Embora a implementação atual foque na coleta, normalização e armazenamento, o esquema relacional padronizado facilita a geração de subconjuntos de dados em formato CSV ou JSON. Esse esquema também facilita a conexão direta dos dados com *pipelines* de aprendizado de máquina e ferramentas de visualização, sem a necessidade de reescrever a lógica de coleta para cada nova utilização.

5 Análise de dados

A partir da ferramenta descrita na Seção 4, é possível construir um conjunto de dados unificado com informações de usuários, mídias e comentários provenientes do Instagram, TikTok, Twitter e YouTube. Cada plataforma alimenta seu respectivo banco SQLite, mas todos seguem o mesmo padrão lógico de esquema, permitindo consultas homogêneas para diferentes tipos de análise. Em um cenário de uso típico, o pesquisador pode, por exemplo, selecionar um conjunto de perfis, coletar suas publicações e comentários em um intervalo temporal definido. Em seguida, é possível utilizar diretamente as tabelas de `user_info`, `media_info` e `comments` para alimentar soluções específicas de análise ou ferramentas estatísticas para explorar padrões de engajamento, frequência de postagem, coocorrência de termos, entre outros indicadores.

Do ponto de vista prático, a existência desse banco padronizado simplifica significativamente o fluxo de trabalho em comparação com ferramentas existentes discutidas na Seção 2 (APIs oficiais, serviços pagos e bibliotecas de *scraping* isoladas). Enquanto boa parte dessas soluções exige um esforço adicional de transformação, limpeza e junção de arquivos heterogêneos (CSV, JSON, etc.), a ferramenta aqui proposta já entrega os dados em um formato diretamente consultável, com chaves consistentes para vincular usuários, mídias e comentários. Isso se reflete tanto na simplicidade de uso (notebooks de análise podem ser reaproveitados entre plataformas mudando somente o prefixo das tabelas) quanto na reprodutibilidade, uma vez que todo o processo de coleta, normalização e persistência é automatizado e documentado no próprio código.

Com essa infraestrutura, uma variedade de análises se torna viável. No contexto de estudos sobre fraudes e comportamento abusivo em mídias sociais, o pesquisador pode filtrar publicações e comparar seus padrões de engajamento com conteúdo legítimo, observando diferenças em volume de comentários, tipos de reações ou vocabulário utilizado nas legendas e respostas. De maneira análoga, a ferramenta permite investigações mais amplas sobre engajamento (distribuição de curtidas, respostas e compartilhamentos por tipo de conteúdo), identificação de perfis potencialmente suspeitos (contas com crescimento atípico, alta repetição de mensagem, uso intenso de links externos) e estudo de dinâmica de discussões em torno de temas sensíveis, considerando simultaneamente texto, mídias e metadados.

5.1 Análise de Execução

Para validar a execução da ferramenta proposta, foi planejado e realizado um conjunto de testes voltados à avaliação da eficiência na coleta de dados. Nesses testes, foram coletadas publicações de diferentes usuários nas quatro plataformas de redes sociais suportadas pela implementação. No experimento, considerou-se a coleta de 30 publicações de cada rede social, contemplando variações no

volume de comentários, curtidas e outras métricas relevantes para a análise de eficiência.

As análises de eficiência consideradas neste trabalho baseiam-se em duas métricas: Taxa de Aquisição de Dados por Segundo (kB/s) e Taxa de Aquisição de Comentários por Segundo (c/s). A primeira métrica fornece uma visão geral do volume de dados transferido e do tempo necessário para recuperar todas as informações associadas às publicações. Já a segunda é particularmente relevante porque, em testes preliminares, observou-se que a maior parte do tempo de coleta é concentrada na recuperação de comentários e, para cada comentário, na obtenção dos dados gerais que caracterizam seus autores.

A Tabela 1 apresenta os resultados consolidados. As medições indicam que, na plataforma YouTube, a ferramenta obteve desempenho significativamente superior, registrando taxas aproximadamente 20 vezes maiores que as demais plataformas na coleta de comentários e até 50 vezes maiores na coleta de dados em geral. Em contraste, o Instagram apresentou o menor desempenho entre as plataformas avaliadas.

A diferença de desempenho observada entre as plataformas está diretamente relacionada às distintas formas de acesso aos dados adotadas em cada uma delas. O YouTube apresentou o melhor resultado, uma vez que sua coleta é realizada por meio da API oficial, a qual é estável e otimizada para lidar com grandes volumes de dados. Já o Instagram e o Twitter/X apresentaram desempenho inferior devido aos mecanismos de detecção de automação empregados por essas plataformas.

No caso específico do Twitter/X, a plataforma aplica bloqueios temporários após determinado volume de acessos. No Instagram, além da redução da taxa de coleta, esses mecanismos de segurança geram alertas frequentes de detecção de atividade automatizada, o que pode levar à desativação da conta utilizada para coleta e exigir a criação de novos perfis para continuidade do processo. Por fim, o TikTok apresentou uma taxa de coleta de comentários satisfatória, aproximadamente 20 vezes superior à observada no Instagram e no Twitter/X, mas manteve uma taxa geral de coleta semelhante as dessas plataformas, devido às estratégias, como pausas temporais, implementadas para evitar o bloqueio das contas coletoras, as quais se mostraram eficazes durante todos os testes e em coletas posteriores.

Em resumo, cada plataforma apresentou desafios específicos no processo de coleta. No YouTube, apesar do bom desempenho, a API impõe um limite diário de dez mil requisições, o que pode ser insuficiente dependendo da quantidade de publicações a serem coletadas. Para mitigar essa limitação, a ferramenta proposta permite o uso de múltiplas chaves de API. No Instagram, a utilização de múltiplas contas reduz o impacto das restrições impostas pelo sistema de detecção de automação. No TikTok, a biblioteca `TikTokApi` apresenta um erro interno que limita o acesso a dados de perfis, já reportado ao projeto oficial. Além disso, por depender do `Selenium`, alterações na interface podem ocasionar a perda de alguns dados. Situação semelhante ocorre no Twitter/X, que depende exclusivamente do `Selenium`, tornando o processo sensível às frequentes mudanças na interface da plataforma.

5.2 Discussão

A ferramenta proposta funciona, na prática, como uma infraestrutura de pesquisa para estudos empíricos em mídias sociais, reduzindo a barreira de entrada entre o interesse em um fenômeno e a disponibilidade de dados adequados para análise. Ao abstrair detalhes específicos de cada plataforma em um esquema relacional comum, o sistema permite que o pesquisador foque nas questões substantivas do estudo, sem precisar reimplementar a cada projeto o mesmo conjunto de rotinas de coleta, limpeza e integração. Essa camada de abstração é particularmente relevante em pesquisas que exigem comparação entre plataformas (por exemplo, diferença na circulação de campanhas fraudulentas no Instagram e TikTok, ou padrões distintos de engajamento no Twitter e YouTube), por garantir que as consultas sejam expressas em termos de entidades conceitualmente equivalentes.

O formato padronizado do banco de dados também facilita a aplicação de métodos de análise multimodal e técnicas de NLP (*Natural Language Processing*). Como textos de legendas, descrições e comentários são armazenados em um campo bem definido, torna-se simples extrair conjuntos para tarefas de modelagem de tópicos, análise de sentimento ou identificação automática de padrões linguísticos associados a fraudes, discurso de ódio ou desinformação. Da mesma forma, o vínculo explícito entre textos, mídias (imagem, vídeos e áudios) e metadados possibilita estudos que combinem, em uma mesma unidade analítica, características visuais, conteúdo textual e sinais de engajamento, abrindo para abordagens híbridas com redes neurais multimodais ou modelos hierárquicos de análise.

Além disso, o uso de bancos relacionais por plataforma, com convenções de nomeação consistente, favorece análises longitudinais e estudos sociotécnicos. Uma vez estabelecido o conjunto de perfis de interesse, é possível repetir a coleta em diferentes momentos do tempo, consolidando séries históricas de métricas e conteúdos que podem ser comparados, por exemplo, antes e após mudanças de políticas de moderação, de grandes eventos sociais ou de campanhas coordenadas. A mesma infraestrutura pode ser reutilizada em múltiplos projetos, contribuindo para a construção de repositórios duradouros de dados e para a replicação de resultados por outros grupos de pesquisa.

Por outro lado, a ferramenta herda limitações estruturais inerentes ao domínio de mídias sociais. Em primeiro lugar, o acesso aos dados permanece condicionado aos termos de serviço de cada plataforma, que pode restringir o tipo de conteúdo obtido, a frequência de requisições ou a forma de armazenamento e redistribuição dos dados. Em segundo lugar, mudanças nas APIs, nas interfaces web ou em mecanismos de proteção contra automação podem introduzir quebras súbitas na coleta, exigindo manutenção contínua dos módulos de *scraping* e das rotinas de autenticação.

Por fim, falhas ocasionais de coleta (como interrupções de conexão, respostas incompletas ou alterações de *layout* não detectadas) podem resultar em amostras parcialmente incompletas, demandando cuidado na interpretação dos resultados e na documentação das condições sob as quais cada conjunto de dados foi obtido. Mesmo com essas restrições, a padronização proposta e a separação clara entre coleta e análise contribuem para tornar esses limites mais visíveis, controláveis e comunicáveis no contexto de pesquisas acadêmicas.

Métrica/Rede Social	**Instagram**	**TikTok**	**Twitter/X**	**YouTube**
Taxa de Coleta de Dados (Postagens)	0.01 kB/s	0.02 kB/s	0.02 kB/s	1.02 kB/s
Taxa de Coleta de Comentários	0.02 c/s	0.19 c/s	0.07 c/s	3.78 c/s

Tabela 1: Dados de Eficiência em Coleta de Postagens

6 Considerações Éticas

A coleta e o tratamento de dados provenientes de plataformas de mídias sociais exige atenção especial a princípios éticos, legais e de responsabilidade no uso da informação. Embora este trabalho proponha uma ferramenta voltada à pesquisa acadêmica, sua operação envolve necessariamente o acesso a conteúdos publicados por usuários, o que demanda cuidado quanto à privacidade, ao contexto de uso e à forma como esses dados podem ser posteriormente analisado e compartilhado.

Primeiramente, a ferramenta foi projetada para atual exclusivamente sobre conteúdo público, acessível sem violação de barreiras técnicas, de autenticação avançada ou de permissões explicitas. Nenhum mecanismo de contorno de segurança, engenharia reversa de APIs privadas ou coleta de informações sensíveis é utilizado. Os dados capturados refletem somente aquilo que já está exposto publicamente pelas plataformas, respeitando a expectativa razoável de visibilidade estabelecida pelos próprios usuários no momento da publicação.

Além disso, o *pipeline* não realiza nenhum tipo de inferência automatizada sobre atributos sensíveis (como orientação política, saúde, identidade ética ou outros aspectos protegidos). O sistema apenas coleta e estrutura os elementos multimodais disponibilizados pelas redes sociais, sem promover classificações que possam gerar riscos adicionais aos usuários analisados.

Do ponto de vista de armazenamento, a ferramenta mantém todos os dados localmente, em bancos de dados independentes por plataforma, sem envio para serviços externos ou infraestruturas de terceiros. Pesquisadores são orientados a adotar práticas seguras de armazenamento e descarte, evitando qualquer redistribuição de conteúdo que possa violar os termos de serviço das plataformas ou comprometer a privacidade dos indivíduos envolvidos. Logs internos, por sua vez, são limitados a informações técnicas indispensáveis ao funcionamento da ferramenta, sem registrar conteúdos de usuários.

Por fim, reforça-se que o objetivo central da ferramenta é reduzir assimetrias de acesso a dados em um cenário de crescente restrição e opacidade das plataformas, oferecendo suporte metodológico para estudos de interesse publico, incluindo investigações sobre segurança, golpes e comportamento coletivo. O uso ético e responsável permanece condição fundamental para garantir que sua contribuição à pesquisa não resulte em danos a usuários ou violação de seu direito à privacidade, observando especialmente legislações vigentes, como a Lei Geral de Proteção de Dados (LGPD), no Brasil, e a *General Data Protection Regulation* (GDPR).

7 Conclusão

Esse trabalho apresentou um *pipeline* de processos e uma ferramenta para coleta e armazenamento de dados multiplataforma em redes sociais, contemplando, inicialmente, Instagram, TikTok, Twitter/X e YouTube. Partindo de um cenário marcado por acesso restrito a dados, limitações de APIs oficiais e fragilidade de soluções baseadas em *scraping* isolado, a proposta oferece uma infraestrutura unificada capaz de coletar, normalizar e persistir, de forma estruturada, informações sobre usuários, mídias e comentários. Ao adotar um esquema relacional padronizado e módulos de coleta específicos por plataforma, a ferramenta reduz o esforço de engenharia necessário para estudos empíricos em mídias sociais e favorece a reprodutibilidade de experimentos.

Do ponto de vista técnico-metodológico, a principal contribuição está na combinação de coleta multimodal (texto, imagens, vídeos, áudios, metadados e URLs) com um modelo de dados coerente entre plataformas. Essa padronização permite que consultas e análises sejam formuladas em termos de entidades conceitualmente equivalentes, abrindo espaço para estudos comparativos entre redes, análises longitudinais e aplicações de métodos de NLP e de análise multimodal. Além disso, a integração de mecanismos de controle de tempo, registro de *logs* e manejo de múltiplas contas e *cookies* contribui para tornar o processo de coleta mais robusto frente a bloqueios, limites de acesso e mudanças incrementais nas plataformas.

Como desdobramento, trabalhos futuros podem ampliar a ferramenta em três direções principais. Primeiro, incorporar novas plataformas e formatos de dados, acompanhando a evolução do ecossistema de mídias sociais e de seus modelos de acesso. Segundo, desenvolver módulos de coleta incremental e agendada, permitindo atualizar periodicamente séries históricas de perfis e campanhas sem repetir todo o processo desde o início. Terceiro, integrar diretamente ao *pipeline*, componentes de análise, como rotinas de pré-processamento textual, detecção automática de fraudes e geração de indicadores de engajamento, de modo a aproximar ainda mais a camada de coleta das necessidades analíticas de diferentes comunidades de pesquisa. Dessa forma, a ferramenta pode consolidar-se não apenas como um coletor de dados, mas como uma peça integrada à infraestrutura tecnológica subjacente que habilita a execução de estudos sociotécnicos em larga escala em mídias sociais.

Agradecimentos

Este trabalho foi realizado no contexto do projeto intitulado Entre Likes e Golpes: Detecção e Alerta de Atividades Fraudulentas em Redes Sociais (Edital UFPR COFPI/PRPI 19/2025), com apoio parcial da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES Brasil), Centro de Computação Científica e Software Livre (C3SL) e Fundação de Amparo à Pesquisa e Inovação de Santa Catarina (FAPESC). Os autores também agradecem o Programa de Pós-Graduação em Informática da Universidade Federal do Paraná.

Referências

[1] Ralph Schroeder. Big data and the brave new world of social media research. *Big Data & Society*, 1(2):2053951714563194, 2014.

[2] Norjihan Abdul Ghani, Suraya Hamid, Ibrahim Abaker Targio Hashem, and Ejaz Ahmed. Social media big data analytics: A survey. *Computers in Human behavior*, 101:417–428, 2019.
[3] Anderson Frasão, Tiago Heinrich, and Vinicius Fulber Garcia. O cenário atual de golpes em redes sociais: Uma revisão da literatura. *Anais do Computer on the Beach*, 16:406–413, 2025.
[4] Roger S Debreceny, Tawei Wang, and Mi Zhou. Research in social media: Data sources and methodologies. *Journal of Information Systems*, 33(1):1–28, 2019.
[5] Kayo Mimizuka, Megan A Brown, Kai-Cheng Yang, and Josephine Lukito. Post-post-api age: Studying digital platforms in scant data access times. *arXiv preprint arXiv:2505.09877*, 2025.
[6] Amelia Acker and Adam Kreisberg. Social media data archives in an api-driven world. *Archival Science*, 20(2):105–123, 2020.
[7] Carlos Entrena-Serrano, Martin Degeling, Salvatore Romano, and Raziye Buse Çetin. Tiktok's research api: Problems without explanations. *arXiv preprint arXiv:2506.09746*, 2025.
[8] Phillip Brooker, Julie Barnett, Timothy Cribbin, and Sanjay Sharma. Have we even solved the first 'big data challenge?'practical issues concerning data collection and visual representation for social media analytics. In *Digital methods for social science: An interdisciplinary guide to research innovation*, pages 34–50. Springer, 2016.
[9] Anuradha Goswami and Ajey Kumar. Challenges in the analysis of online social networks: A data collection tool perspective. *Wireless Personal Communications*, 97(3):4015–4061, 2017.
[10] Stefan Stieglitz, Milad Mirbabaie, Björn Ross, and Christoph Neuberger. Social media analytics–challenges in topic discovery, data collection, and data preparation. *International journal of information management*, 39:156–168, 2018.
[11] Katrin Weller and Katharina Kinder-Kurlanda. Uncovering the challenges in collection, sharing and documentation: the hidden data of social media research? In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 9, pages 28–37, 2015.
[12] Make. Ai automation you can visually build and orchestrate in real time, nov 2025. https://www.make.com/en.
[13] bright data. The web's data, unlocked, nov 2025. https://brightdata.com/.
[14] Thunderbit. The next gen ai web scraper, nov 2025. https://thunderbit.com/.
[15] apify. Get real-time web data for your ai, nov 2025. https://apify.com/.
[16] Octoparse. Easy web scraping for anyone, nov 2025. https://www.octoparse.com/.
[17] TikTok for developers. Develop for >communities, nov 2025. https://developers.tiktok.com/.
[18] X. X api, nov 2025. https://developer.x.com/en/docs/x-api.
[19] Instaloader. Instaloader, nov 2025. https://instaloader.github.io/.
[20] snscrape. snscrape, jun 2023. https://github.com/JustAnotherArchivist/snscrape.
[21] yt dlp. Yt-dlp a feature-rich command-line audio/video downloader, nov 2025. https://github.com/yt-dlp/yt-dlp.
[22] Ismael Camargo-Henriquez and Yarisel Nunez-Bernal. A web scraping based approach for data research through social media: An instagram case. In *2022 V Congreso Internacional en Inteligencia Ambiental, Ingeniería de Software y Salud Electrónica y Móvil (AmITIC)*, pages 1–4. IEEE, 2022.
[23] Lucía Martín-Gómez, Rebeca Cordero-Gutiérrez, and Javier Pérez-Marcos. Business benefits of instagram scraping: Questionable uses of data. In *International Conference on Disruptive Technologies, Tech Ethics and Artificial Intelligence*, pages 219–227. Springer, 2021.
[24] Stefanos Vlachos, Dimitris Linarakis, Nikos Platis, and Paraskevi Raftopoulou. Exploring instagram and youtube data. In *Data Management Technologies and Applications: 10th International Conference, DATA 2021, Virtual Event, July 6–8, 2021, and 11th International Conference, DATA 2022, Lisbon, Portugal, July 11-13, 2022, Revised Selected Papers*, page 90. Springer Nature, 2023.
[25] Johannes B Gruber. traktok—making tiktok data accesible for research. 2025.
[26] Yannick Zelle, Thibault Grison, and Marc Feger. Scitok-a web scraping tool for social science research. In *International Conference on Human-Computer Interaction*, pages 103–109. Springer, 2023.
[27] Jaebeom You, Jaekyu Lee, and Hyuk-Yoon Kwon. A complete and fast scraping method for collecting tweets. In *2021 IEEE international conference on big data and smart computing (BigComp)*, pages 24–27. IEEE, 2021.
[28] Parneet Kaur. Sentiment analysis using web scraping for live news data with machine learning algorithms. *Materials today: proceedings*, 65:3333–3341, 2022.
[29] Joseph Kready, Shishila Awung Shimray, Muhammad Nihal Hussain, and Nitin Agarwal. Youtube data collection using parallel processing. In *2020 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, pages 1119–1122. IEEE, 2020.
[30] Rachel E Scott. Data scraping youtube for the study of lieder reception. *Nineteenth-Century Music Review*, 19(3):655–667, 2022.