Data Donation on Social Media: Tools and Datasets

Kiran Garimella

kiran.garimella@rutgers.edu

Abstract

Access to social media data is becoming increasingly restricted as platforms tighten their policies, posing significant challenges for research. End-user-supported data donation offers a promising solution, yet the practical implementation of such approaches remains under explored. In this paper, we present a suite of tools for data donation that we developed and deployed across five major social media platforms: WhatsApp, Facebook, Telegram, YouTube, and Instagram. These tools were designed with scalability and usability in mind, enabling the participation of hundreds of users.

We also release several novel datasets collected using these tools, including viral WhatsApp messages from three countries, complete YouTube watch histories from hundreds of American users, and data collected via a mobile application for Instagram. These contributions mark significant advancements in the ability to collect, analyze, and understand user-specific social media data at scale. By sharing these tools and datasets, we aim to inspire further research into data donation methodologies and to support the broader academic community in navigating the increasingly constrained landscape of social media research.[1]

1 Introduction

The availability of social media data for research and academic inquiry is rapidly declining due to increasing restrictions by major platforms (Lukito 2024). These restrictions pose significant challenges to understanding the societal, cultural, and political dimensions of social media, which have become deeply intertwined with everyday life (Lazer et al. 2021). However, reliance on platform-controlled APIs or third-party aggregators leaves researchers at the mercy of shifting policies and opaque restrictions.

To address these challenges, this paper explores end-user-driven data donation as a viable alternative. By empowering individuals to voluntarily share their social media data, this approach circumvents the need for platform cooperation and enhances data accessibility. Specifically, we introduce a suite of tools designed for five major platforms –WhatsApp, Facebook, YouTube, Telegram, and Instagram– and present datasets collected through these tools. These tools are tailored to make data donation user-friendly, privacy-compliant, and accessible across a variety of platforms and devices. The datasets resulting from this approach are rich in detail and offer unprecedented opportunities for diverse types of social media research.

Data donation methods often complement traditional data collection approaches via official APIs, offering distinct advantages that enhance research capabilities. As detailed in (van Driel, Giachanou et al. 2022), these advantages include the ability to achieve more representative sampling, collect data longitudinally, reduce costs, and work effectively with smaller sample sizes. The tools we present aim to capitalize on these benefits while addressing the limitations and challenges inherent in previous methods.

Given the critical role that social media plays in shaping public discourse, developing scalable and standardized data donation approaches is of paramount importance. Platform policies are subject to change without notice, jeopardizing research projects and creating substantial uncertainties for longitudinal analyses. Researchers are left to grapple with disruptions in data availability that hamper the study of key phenomena, such as misinformation, mental health trends, and large-scale social mobilization. While data donation models have been proposed in the past (Araujo et al. 2022; Boeschoten et al. 2022), they often remain fragmented, lack consistent standards, and fail to address the expanding diversity of social media platforms and user behaviors. Strengthening the infrastructure for data donation, particularly by creating robust software tools and protocols, can unify these efforts and catalyze more nuanced, ethical, and reproducible social media research. This paper takes a step in that direction by introducing tools designed with scalability, accessibility, and ease of use in mind.

Realizing a user-centric data donation system that works across platforms, devices, and operating systems is far from trivial. The naive approach of simply requesting users to download and hand over data exports quickly becomes infeasible due to platform-specific data structures and file formats that may be incomplete or poorly organized for research purposes (van Driel, Giachanou et al. 2022; Hase et al. 2024). Moreover, current solutions often rely on desktop-based browser extensions that are not readily accessible to the significant number of users who engage with

[1]The datasets from thiis paper are available at https://doi.org/10.7910/DVN/VOFPK1. Refer to Table 1 for code.

social media exclusively on mobile devices (Boeschoten, de Schipper et al. 2023). Ensuring that such tools can operate without excessive computational overhead –and within realistic budgets– poses additional challenges. It is also insufficient merely to collect data; developers must factor in the ethical obligations of privacy, consent, and the security of sensitive user information.

In addition to the tools, we release several valuable and novel datasets that hold significant importance for the research community. These datasets include viral content shared on WhatsApp across multiple countries, lists of public groups, pages, and channels subscribed to by users on Facebook and Telegram, complete YouTube watch histories, and a snapshot of user feeds on Instagram.

We hope this paper serves as a meaningful contribution to research in this domain, providing tools and methodologies that can be replicated, extended, or adapted to other platforms. This work is particularly aimed at supporting disciplines like the social sciences, where technical expertise for such endeavors may be less readily available, fostering broader engagement and innovation in social media research.

Ethics note: All data collection described in this paper was conducted following a thorough consent process, ensuring that participants were fully informed about the data being collected and its intended use. No publicly identifiable information is being released, and out of abundance of caution we restrict access to some of the datasets for academic use only. The tools and datasets we release have been reviewed and approved by the Rutgers Institutional Review Board under protocol numbers Pro2022000312, Pro2022001023, Pro2023001387, and Pro2024002400.

2 Related work

Data donation has increasingly emerged as a compelling solution to the challenges posed by shifting platform policies and dwindling levels of cooperation from social media companies. Scholars have conceptualized data donation as an exercise in personal sovereignty that can forge new forms of social bonding and recognition (Hummel, Braun, and Dabrock 2019). This approach has gained renewed urgency with platforms like Twitter drastically reducing or revoking public API access, thereby inhibiting many long-term research projects (Lukito 2024).

Nevertheless, while data donation helps researchers bypass some of the barriers created by platform lockdowns, it also raises critical ethical considerations. Users who voluntarily donate their social media traces may inadvertently expose private information about friends and acquaintances, highlighting the complexity of managing third-party data in donation-based frameworks (Boeschoten et al. 2022; Gomez Ortega et al. 2023; Garimella and Chauchard 2024). Moreover, widespread uncertainty persists around what users truly consent to when donating, as many remain unaware of how their data is collected, stored, or repurposed.

Although certain legal frameworks, particularly in the European Union, nominally grant users the right to download comprehensive archives of their activity under the General Data Protection Regulation (GDPR), these exports often fall short of research needs (van Driel, Giachanou et al. 2022). YouTube data, for instance, can be downloaded, but the structures and metadata are neither standardized nor consistently maintained. Facebook's export provides even fewer actionable identifiers, omitting page or group IDs that would enable deeper analyses of user interactions beyond the personal profile. Researchers are further impeded by platform noncompliance, as documented in (Hase et al. 2024) that nominal data-access laws are not fully respected by many services, resulting in incomplete or obscured data.

Efforts to compensate for these shortcomings with open-source frameworks strive to standardize, annotate, and unify donated data across platforms (Araujo et al. 2022), while others emphasize user-centric pipelines for particular demographics, as in the case of adolescent Instagram donors (Razi, AlSoubai et al. 2022). Even with explicit user permission, however, access can be abruptly terminated or restricted by platforms, as illustrated by incidents involving researchers from New York University (Bond 2021), showcasing the tenuous legality of Terms of Service agreements and their frequent ambiguity (Fiesler, Beard, and Keegan 2020).

In response to these obstacles, several large-scale observatories and research consortia have mobilized to promote data donation as a robust, community-driven strategy. The National Internet Observatory, proposed by (Feal et al. 2024), envisions an infrastructural ecosystem to gather and analyze digital communication data in a more transparent manner. Parallel initiatives have been created in Australia (Angus et al. 2024), Germany (Leibniz Institute for Media Research 2024), and other contexts (dat 2024; uzh 2024; Norton and Shapiro 2024). Collectively, these initiatives are a welcome progress in a global push toward open, ethical, and sustainable infrastructures for studying online behavior while addressing key operational, legal, and technological challenges.

Our contribution to the field of social media data donation is both unique and innovative in several key ways. Unlike many existing approaches that rely primarily on platform-provided data exports, our work emphasizes the development of tools that leverage novel platform features to operationalize data donation at scale. Our goal is to create reusable and adaptable tools that can support diverse research needs, rather than simply producing another standalone dataset.

3 Tools for social media data donation

To add to the literature, we have developed a suite of tools that go beyond the standard data exports and tap into the unique features and functionalities of each platform. By doing so, we aim to unlock new avenues for data collection and analysis that were previously inaccessible or difficult to achieve at scale.

Our tools are designed with user-friendliness, cross-platform compatibility, and mobile accessibility in mind. We prioritize the development of intuitive interfaces and clear documentation to facilitate easy adoption by researchers and end-users alike. By making the data donation process more accessible and streamlined, we hope to encourage greater participation and enable a scalable data donation system.

In the rest of the section, we will briefly discuss the tools we developed for 5 major platforms – WhatsApp, Facebook, YouTube, Telegram, and Instagram. Table 1 provides a detailed description of our tools, including links to the code, a deployed web instance, and a video walkthrough demonstrating the tool's functionality.

3.1 WhatsApp

WhatsApp, with over 2 billion monthly active users, stands as one of the most popular social networks globally, particularly in the Global South. Its encrypted, chat-based format makes it a significant platform for understanding social interactions, yet it also presents numerous challenges for data collection due to its private and secure communication design.

Our tool, WhatsApp Explorer (Garimella and Chauchard 2024), is specifically designed to facilitate the collection of quantitative data from WhatsApp while addressing the unique ethical, legal, and practical challenges associated with this platform. Given WhatsApp's end-to-end encryption, the only viable method for data collection is through user-driven data donation. This strategy necessitates careful consideration to ensure ease of use, safeguard donor privacy, foster trust among diverse participants, and minimize legal risks for the research team.

To meet these requirements, we developed a specialized web interface for WhatsApp Explorer. This tool is designed to address the logistical and ethical challenges of collecting data from private WhatsApp groups while enabling the large-scale acquisition of data for research purposes.

WhatsApp Explorer leverages the capabilities of whatsapp-web.js, an open-source library that functions as a WhatsApp client library for Node.js. Node.js, a JavaScript runtime, allows server-side execution of JavaScript, enabling programmatic interaction with WhatsApp through its web browser application. The whatsapp-web.js library operates by reverse-engineering API calls used by WhatsApp Web, enabling the tool to authenticate, read, and process messages from the user's account in an automated yet secure manner.

To initiate the process, users authenticate themselves by scanning a QR code through the WhatsApp Web interface. Once authenticated, users select specific groups they wish to donate data from and answer a series of preliminary questions designed to understand their consent and data-sharing preferences. Using official WhatsApp APIs, the tool queries and retrieves content from the specified groups. Our protocol restricts data collection to a defined timeframe: two months prior to and two months after the user's recruitment. Users retain full control over the process, with the ability to disconnect their account and cease participation at any time. To protect user privacy, we immediately anonymize all collected data. Personal identifiers such as names, email addresses, and phone numbers are replaced with unique, non-identifiable codes. Additionally, sensitive visual data such as images containing faces are blurred before the content is stored on secure servers.

For a detailed account of the privacy measures, ethical considerations, and the rationale behind our design choices, we refer readers to the discussion in (Garimella and Chauchard 2024). The code for WhatsApp Explorer is available for academic use (upon request) at https://github.com/gvrkiran/WhatsAppExplorer.

3.2 Facebook

Facebook remains one of the most widely used social media platforms globally, with over 2.9 billion monthly active users as of 2023, making it a vital site for understanding online behavior, social interactions, and information dissemination. Its extensive reach across diverse demographics and regions has established Facebook as a critical platform for news consumption, political discourse, and community engagement. Studies have shown that Facebook is among the primary sources of news for millions worldwide, with significant implications for public opinion and democratic processes (Walker and Eva Matsa 2021). Additionally, Facebook's unique features, such as public groups and pages, foster public spaces where specialized communities and facilitate the spread of information, both accurate and misleading. This dual role as a space for both connection and potential misinformation makes it an important platform to study in the context of understanding the dynamics of digital communication and media ecosystems.

Our approach to retrieving Facebook data enables users to donate their lists of public Facebook pages and groups to which they belong. Using these lists, we can leverage Facebook-provided tools such as CrowdTangle or the Meta Content Library to retrieve content from these pages and groups. While the algorithmic nature of Facebook's feed means there is no guarantee that users have seen all the content from these pages or groups, this collection serves as an upper bound of potential content exposure based on their chosen memberships. To facilitate this process, we developed a Facebook app similar to apps like Farmville (Burroughs 2014). This app allows users to log in with their Facebook credentials and selectively grant access to the groups and pages they are comfortable donating. The application uses the Facebook Graph API, focusing on specific endpoints such as user_likes and groups_access_member_info to retrieve the lists of pages and groups. This approach ensures a user-centric and flexible method for data donation, supporting ethical research on social media content exposure. Our application underwent Facebook's rigorous manual approval process to justify each requested permission. Permissions are scoped to only those necessary for research purposes, and users retain control over what data they wish to share and can request to delete their data.

Our tool is built using Django, a high-level Python web framework. Once users log in and consent to share their data, the application collects only the list of groups and pages they specify. The collected information includes group and page IDs, which serve as inputs for obtaining content from these sources. For this purpose, we utilize Meta's research tools, including CrowdTangle and the Meta Content Library[2] (formerly known as Facebook Open Research and Transparency

[2]https://transparency.meta.com/en-gb/researchtools/meta-content-library/

Table 1: Details of the tools developed for various platforms. While the deployed instance may not remain available indefinitely, the code and walkthrough provide valuable resources for understanding and utilizing the tools in the future.

Platform	Code	Deployed instance	Walk through
WhatsApp	https://github.com/gvrkiran/WhatsAppExplorer	https://whatsapp.whats-viral.me/	https://youtu.be/_NGIJG4a-hY
Facebook	https://github.com/gvrkiran/facebook-data-donation	https://diaspora-watch.us/	https://youtu.be/VQrgxcgMdEQ
YouTube	https://github.com/gvrkiran/google-data-donation	https://data-donation.vercel.app/	https://youtu.be/ytsunBxCWW8
Telegram	https://github.com/gvrkiran/telegram-data-donation	https://telegram.whats-viral.me/	https://youtu.be/vGih8FJ4TwE
Instagram	https://github.com/gvrkiran/instagram-data-donation	https://play.google.com/store/apps/details?id=com.rutgers.smdr	https://youtu.be/FE_LMO-AOnQ

(FORT)). To align with Facebook's data access policies, our tool is limited to collecting content from large, public groups and pages with more than 25,000 followers. For details on the data collection design, flow, and validation, as well as broader discussions on ethical considerations, we refer readers to (Couto and Garimella 2024).

3.3 YouTube

YouTube stands as one of the most influential social media platforms globally, with over 2.5 billion active users, making it second only to Facebook in terms of user base. Its extensive reach and diverse content have transformed it into a central hub for news consumption, entertainment, and education. Notably, a significant portion of users turn to YouTube for news, with both established news organizations and independent creators contributing to the platform's rich information ecosystem. The platform is significantly popular across various demographics, serving as a primary source of information and influence, particularly among younger audiences (Stocking, Galen and van Kessel, and others 2020).

To address the challenges of obtaining YouTube usage data for research, we developed a Google application designed to facilitate the seamless donation of user data through Google Takeout.[3] Traditional methods for accessing YouTube data, such as direct downloads via Takeout, are plagued by significant inefficiencies. The Takeout process is unpredictable, with archive generation times varying from a few minutes to several days, often leaving users uncertain about when their data will be available. This uncertainty, combined with the manual effort required to download large zip files and share them with researchers, makes participation particularly inconvenient, especially for users relying on mobile devices. These barriers significantly limit the scalability and accessibility of data donation efforts.

Our application directly integrates with Google Drive, simplifying the process by automating the retrieval of YouTube data stored in Takeout archives. This system utilizes Google OAuth2 authentication to enable users to log in and grant selective access to their Google Drive, using the Google Drive SDK.[4] Once authenticated, the application monitors the user's Google Drive for the availability of the Takeout folder, eliminating the need for users to repeatedly check the status of their export. When the Takeout data becomes available, the system retrieves it and deletes our app's access to the user's Google drive. This approach minimizes user involvement after the initial setup, making the process more scalable and user-friendly.

The flow begins with users initiating a Takeout request at https://takeout.google.com/ to export their YouTube data. They then provide our Google app access to their Google Drive. Our app periodically checks for the presence of the Takeout folder in the user's Drive and automatically processes it once available. User credentials are immediately destroyed once the data is downloaded. Though we are only currently downloading YouTube data using this method, the same methodology can be applied to any other Google property (such as Chrome or location data) which stores data on Google Takeout.

3.4 Telegram

Telegram has experienced significant growth, with its user base expanding from 35 million in March 2014 to approximately 950 million monthly active users by July 2024 (ThinkImpact 2024). This rapid expansion in popularity indicates its global appeal and the platform's pivotal role in information dissemination. Notably, during the Russia-Ukraine conflict, Telegram emerged as a crucial medium for real-time news sharing and coordination among pro-Ukrainian cyber resistance groups, highlighting its importance in contemporary geopolitical events (Canevez, Maikovska, and Zwarun 2024). However, Telegram's commitment to user privacy and minimal content moderation has also made it a haven for extremist groups (Baumgartner et al. 2020).

Telegram Watch, our custom-built tool for Telegram, enables users to ethically donate their list of public channels and groups for research purposes. Similar to our approach for Facebook, this tool facilitates data donation by allowing users to log in through their Telegram accounts and grant

[3] https://takeout.google.com/

[4] https://developers.google.com/drive/api/guides/about-sdk

access to their group and channel lists. This approach addresses one of the most significant challenges in studying Telegram: discovery of content. While there are established methods for collecting public data from Telegram, identifying relevant groups and channels remains a critical bottleneck. Publicly available tools such as https://tgstat.com/ offer partial insights but fail to provide a comprehensive discovery mechanism for groups and channels. For instance, if one were to investigate the channels used by young Americans to consume news about the Russia-Ukraine conflict, the absence of an exhaustive directory of user-specific group memberships makes it nearly impossible to identify and monitor all relevant sources. This gap highlights the importance of user-driven data donation to map and analyze consumption patterns effectively.

Telegram Watch is built with FastAPI and Prisma, leveraging Telegram's SDK via Telethon.[5] The platform allows users to securely authenticate their accounts and donate their lists of groups and channels for research purposes, with explicit consent. Following authentication, the Telegram API is used to programmatically access and retrieve content from the selected sources. The platform enables user authentication through phone numbers and one-time passwords, creating unique Telethon sessions for each user. Users can view and manage their list of Telegram channels, granting permissions selectively for data donation. To enable user privacy and control, the platform includes a data deletion endpoint that allows users to remove their data at any time. The platform also collects user survey data, such as ethnicity, gender, and age, which is linked to the user telegram data.

3.5 Instagram

Instagram has become one of the most influential social media platforms, yet there is a surprising dearth of research tools available for studying its content and user behaviors. Unlike Twitter, which has historically offered extensive public data access, Instagram has remained relatively underexplored, particularly following the discontinuation of tools like CrowdTangle. This lack of research infrastructure is especially concerning given Instagram's significance, not only as a popular platform but also as a space with profound implications for teen mental health (Stefana et al. 2022). Recent changes to Instagram's interface, such as its adoption of a TikTok-style feed (Hutchinson 2021), further emphasize the platform's evolving role in shaping user experiences and its highly personalized nature. However, Instagram's primary use on mobile devices and its dynamic, individualized content streams make it challenging to systematically measure what users are exposed to. To address these gaps, we developed an innovative solution: a custom-built Android app that enables the collection of Instagram consumption data by recording user feeds in real time.

The Android app replicates the Instagram experience by providing users with a mobile web version of the platform. This The experience of using our app closely resembles the native Instagram app in terms of appearance and functionality, with minor differences such as the absence of the ability to upload stories. Preliminary testing suggests that the content served via this web-based interface is largely consistent with that of the native app, though differences in feed order may exist. While further validation is required, these initial observations provide a solid foundation for leveraging the app as a research tool to collect Instagram data.

Users install the our app through the Google Play store and log in with their Instagram credentials, after which they can browse the platform as they would through the native Instagram Android app. During this interaction, the app records all the content the user is viewing. While we only record and release information on the 'Home' feed, the app has functionality to record the 'Explore' and 'Reels' feeds too. Additional features, such as the ability to inject or remove items from the feed, monitor the use of other apps, and alert users when they open the official Instagram app instead of our tool, were also implemented. While these features are not central to the dataset presented in this paper, they offer significant potential for future studies and interventions on Instagram.

The app is a hybrid design, combining native Android components with GeckoView, a robust browser engine developed by Firefox. GeckoView provides advanced capabilities for handling web content, overcoming the limitations of the standard WebView. This flexibility is crucial for enabling detailed interactions with Instagram's web-based functionalities. The app's native Android components handle auxiliary features such as user surveys, permission checks, and detection of other app usage. Meanwhile, the core interactions with Instagram occur through GeckoView, which allows us to extend the app's capabilities via a custom browser extension. This extension overrides Instagram's Content Security Policy, intercepts network requests, manages page loads, and injects custom scripts to interact with the platform. Communication between the native Android code and the GeckoView extension is achieved through GeckoView's messaging APIs. These APIs allow bidirectional communication, enabling the native code to send commands to the extension and receive responses in real time.

By providing a mobile app that records real-world Instagram usage, this tool addresses critical gaps in the study of Instagram's personalized and highly mobile-centered content delivery. It offers researchers a scalable and adaptable solution for studying Instagram feeds, opening new avenues for exploring the platform's impact on user behavior and important factors such as mental health.

4 Datasets

In this section, we describe the datasets that can be collected using our data donation tools. Unlike traditional social media data collection methods, which often rely on keyword-based queries or convenience samples, data donation tools allow us to collect user-level data that can be sampled to meet specific research needs. For most platforms, our tools enable the collection of user-level data on potential exposure, a capability that was previously unavailable for several platforms. This type of data provides insights into the content users could potentially encounter based on their memberships, subscriptions, or interactions, rather than relying

[5] https://docs.telethon.dev/en/stable/

Table 2: Summary of our datasets. The datasets are available at https://doi.org/10.7910/DVN/VOFPK1.

Platform	Data details	Data type	Location
WhatsApp	361 groups, 1,583 users, 7,495 viral messages	Potential exposure	India, Colombia, Indonesia
Facebook	1,336 users, 388k pages/groups	Potential exposure	US
YouTube	285 users, 4.3 million videos, 492k channels	Exposure	US
Telegram	329 users, 12,800 channels/groups	Potential exposure	US
Instagram	99 users, 10,563 posts	Exposure	US

solely on what they actively engage with. Similar efforts to reconstruct potential exposure have been attempted on public platforms like Twitter by simulating user feeds based on follower networks (Eady et al. 2023). Our tools extend this capability to other platforms, offering a more comprehensive view of user-specific content exposure across diverse social media environments.

A summary of the datasets along with their properties is shown in Table 2. All personally identifiable information such as user emails, profile names and phone numbers have been anonymized in the released datasets.

4.1 WhatsApp

We deployed WhatsApp Explorer in six countries: India, Brazil, Colombia, Indonesia, the United States, and Sierra Leone. For this paper, we focus on and release data from three distinct contexts: India, Colombia, and Indonesia. Given the private nature of WhatsApp, we are only releasing anonymized viral content and not participant demographic information.

The dataset includes viral content –messages that have been forwarded multiple times– captured from WhatsApp groups.[6] For each message, we provide the following anonymized metadata: the unique identifier of the sender, the message identifier, the timestamp of the message, and the group identifier.

The data collection process is detailed in (Garimella et al. 2024), which covers the locations of the study: Jharkhand, India; Jakarta, Indonesia; and various regions across Colombia. For further methodological details, we direct readers to that paper. The released dataset includes 2,379 pieces of viral content from India, 364 from Indonesia, and 4,752 from Colombia. This is the first dataset of its kind at such scale for WhatsApp, offering new opportunities for research.

Potential applications of the dataset include studying the prevalence of misinformation, conducting cultural analytics, and examining the dynamics of virality on WhatsApp. By

[6]We consider messages that have been marked as 'Forwarded many times' as viral messages. This designation indicates that a message has been forwarded through a chain of five or more separate users from the original sender.

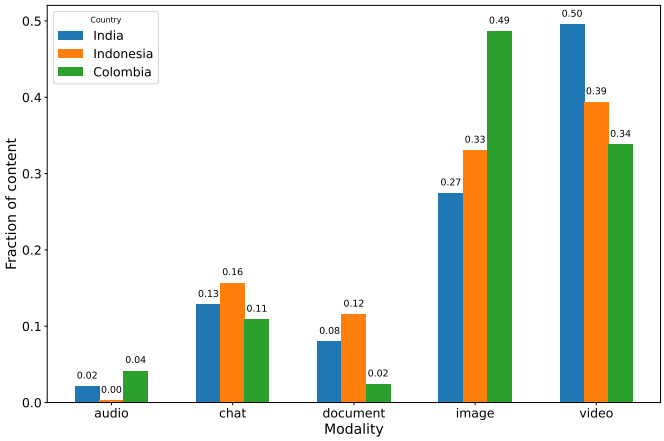

Figure 1: Top 5 modalities in the WhatsApp dataset.

enabling researchers to analyze what types of content go viral and how, this dataset offers valuable insights into one of the most widely used encrypted communication platforms.

Figure 1 illustrates the distribution of content modalities within the dataset. It is notable how the prevalence of different modalities varies across countries. For example, in India, videos account for half of the viral content, whereas in Colombia, they comprise only one-third.

4.2 Facebook

We deployed our Facebook data donation tool on PureSpectrum, a platform known for providing survey panels frequently used in academic research (Lazer, Santillana et al. 2020). Users who participated in the study donated access to their Facebook groups and pages and completed a brief five-question survey about their demographics. The entire process was streamlined, requiring 3–5 minutes per user.

Through this deployment, we collected data from 1,336 users based in the United States. The resulting dataset included a total of 388,890 Facebook pages and groups, comprising 332,427 pages and 56,455 groups. To enhance the dataset, we retrieved posts from a random sample of these pages and groups using CrowdTangle. However, in accordance with CrowdTangle and Meta's privacy policies, we are only able to share the page and group IDs along with anonymized demographic information from the users. Researchers interested in obtaining the associated posts can utilize Meta's academic tools, such as the Meta Content Library to access the data.[7]

Figure 2 shows the distribution of demographics in our dataset. We observe that even though there are differences across different demographic groups, we cover all age groups, genders and ethnicity.

Figure 3 shows the average number of pages and groups per demographic category. We can see clear differences in number of pages/groups people are a part of based on age (older people have significantly lower), gender (men have significantly lower) and ethnicity (whites have significantly

[7]https://transparency.meta.com/en-us/researchtools/meta-content-library/

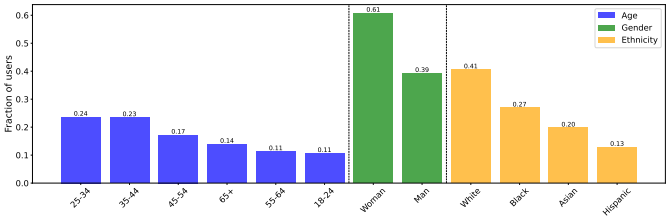

Figure 2: Demographics in our Facebook dataset.

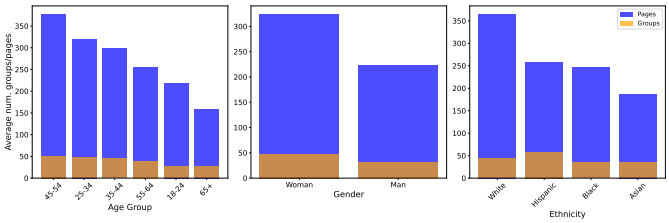

Figure 3: Average number of groups/pages per demographic category.

higher). The figure also shows the fraction of pages and groups. We can see that users are significantly more likely to be part of pages than groups.

The Facebook data we collected offers a wealth of potential applications, particularly in the realm of cultural analytics. By combining user demographics with lists of public pages and groups they have engaged with, this dataset enables researchers to analyze patterns of media consumption and compute potential exposure to various types of content. For instance, this data can be leveraged to explore how different demographic groups interact with political and partisan content over time, as demonstrated by studies such as (Couto and Garimella 2024). Additionally, the dataset provides an opportunity to understand broader behavioral trends, such as how individuals from diverse age groups, geographic locations, or socioeconomic backgrounds allocate their attention across online communities. These insights can inform studies on digital media's role in shaping public opinion, fostering social connections, and perpetuating cultural norms, making this dataset a valuable resource for interdisciplinary research in media studies, sociology, and political science.

Figure 4 shows a word cloud of the groups and pages from the dataset. Some of the dominant topics include expressions of personal identity and community, such as 'Fan,' 'Club,' 'Life,' 'Family,' and 'Home,'; commercial and transactional activities like 'Sale,' 'Buy,' and 'Trade'; cultural and informational content like 'News,' 'Music,' and 'Arts'; local and educational community content like 'County,' 'School,', 'Center,'; and digital subcultures like cooking with keywords like 'Food,' and 'Recipe'. This dataset presents rich opportunities to study social dynamics, community building, and content sharing on Facebook, both among mainstream and long-tail interests.

4.3 YouTube

Our YouTube data donation tools were deployed during the 2024 U.S. elections in October 2024, recruiting 285 Amer-

Figure 4: Word cloud of the page/group titles on Facebook

ican users through Prolific. These participants donated their YouTube watch histories, offering a detailed view of their consumption patterns on the platform. The dataset includes information on all content viewed by users while logged into their Google accounts, spanning both mobile and desktop devices. Notably, the data contains records extending back to the creation of the users' YouTube accounts, with some histories going as far back as 2010, providing a longitudinal perspective on viewing habits over more than 15 years in certain cases.

The dataset includes 4.3 million YouTube videos from 492,000 unique channels, with user histories spanning an average of 1,800 days (median: 1,409 days). It also contains information about advertisements viewed and detailed demographic data on participants, such as political affiliation, gender, age, and ethnicity. The dataset also contains channel-level details such as channel names and can be complemented with additional metadata retrieved through YouTube's API. The YouTube API offers extensive functionality, allowing researchers to gather information such as video titles, descriptions, view counts, and comments.[8]

Given the sensitivity of the data, access is restricted to academic use and available only upon request. Files are hosted under Restricted Access permissions to ensure compliance with ethical standards and participant confidentiality.

Potential applications of this dataset include examining content consumption during the election season, analyzing the effects of YouTube's recommendation system, understanding advertisement targeting, and exploring differences in content consumption across demographic groups. To the best of our knowledge, this represents one of the first large-scale datasets of YouTube viewing histories, offering unprecedented opportunities for research on digital media consumption and its societal impacts.

Figure 5 presents the demographic distribution of users in our sample. It is important to note that this dataset was generated through a convenience sample rather than a representative sampling strategy. As a result, there are observable biases in the demographic composition. For instance, older individuals are underrepresented, and white users are disproportionately overrepresented in the dataset.

Figure 6 illustrates the number of videos in the dataset across various demographic groups. Some notable trends

[8]https://research.youtube/

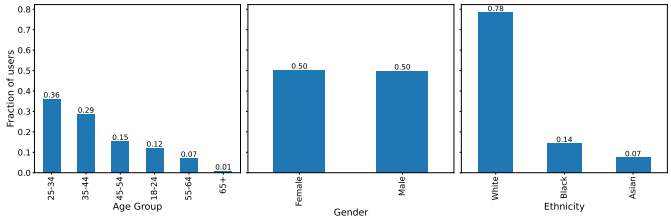

Figure 5: Demographics of YouTube users

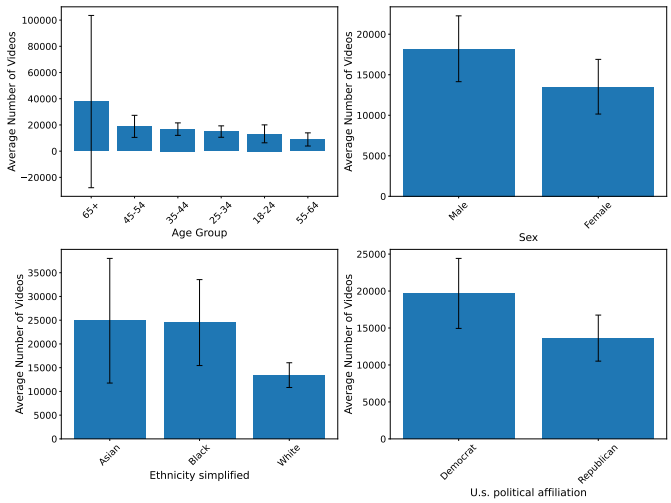

Figure 6: Average number of videos contributed by various demographic groups.

emerge: men and Democrats contribute significantly more videos compared to women and Republicans, respectively. Interestingly, while white users are overrepresented in the dataset, they contribute a significantly lower number of videos per user compared to other demographic groups.

4.4 Telegram

The dataset we collected extends prior Telegram research (Baumgartner et al. 2020) by incorporating user-specific demographic data alongside the list of channels and groups from which participants consume information. In October 2024, we recruited 329 participants on Prolific. Participants donated the lists of Telegram channels and groups to which they were subscribed, resulting in a dataset encompassing 12,800 public channels and groups. Our recruitment specifically targeted Republican users in the United States to explore election conspiracy theories and far-right organizing activities on Telegram, areas of significant scholarly interest given the platform's documented role in such activities (Walther and McCoy 2021).

The channel and group data obtained can be further enriched using existing tools like Telethon, which facilitate the retrieval of content from public Telegram groups and channels. Additionally, we collected detailed demographic information about participants through Prolific, including age, ethnicity, country of birth, country of residence, nationality, primary language, student status, and employment status. These demographic variables are summarized in Figure 7.

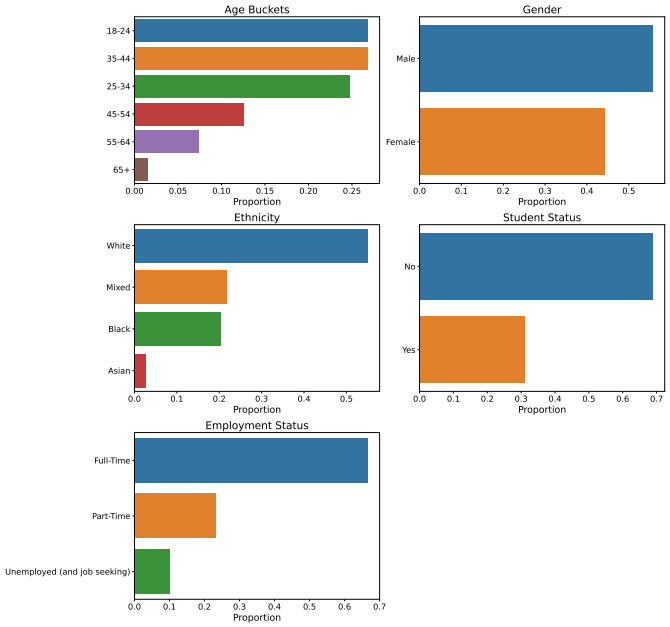

Figure 7: Demographics of telegram users

By combining user demographics with channel subscription data, this dataset offers a unique opportunity to study information consumption patterns, community engagement, and the role of Telegram in contemporary social and political discourse.

Figure 8 shows a word cloud of the telegram dataset. This word cloud highlights key themes from Telegram groups and channels in the dataset. Notably, there is a strong presence of cryptocurrency-related topics, such as 'Crypto,' 'Trading,' 'Airdrop,' 'Wallet,' and 'Token,' showing Telegram's popularity for discussions on blockchain and digital assets (Smuts 2019). Piracy is also a significant theme, with words like 'Movie,' 'Series,' and 'TV' pointing to the sharing of media content. Additional topics include gaming, deals ('Free,' 'Deals'), and educational content ('Tutorial,' 'Academy'). The dataset provides intriguing opportunities to explore Telegram's role in fostering niche communities, enabling decentralized finance discussions, and facilitating media distribution. The dataset has a much broader potential than just these areas. For instance, even though they might not be prominent in the word cloud, the dataset contains dozens of channels relating to the QAnon conspiracy, and discussions on other far right figures which can not be found on other main stream social media platforms.

4.5 Instagram

Among all the datasets presented, the data collected from Instagram is the most preliminary and experimental. We recruited 99 American participants through Prolific to test our app (detailed in Section 3.5). Participants were instructed to log in to their Instagram accounts and use the platform for 15 minutes as they normally would.[9] During this ses-

[9]The 15-minute limit was a practical choice to ensure compliance with the study design. While alternative models for deploying

Figure 8: Word cloud of channel/group names in the Telegram dataset.

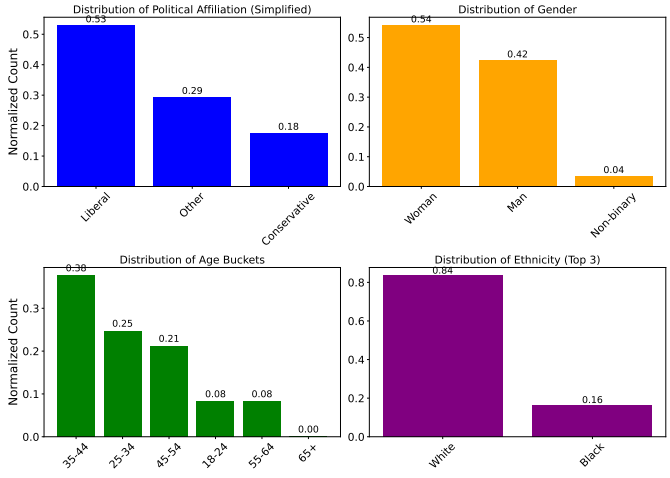

Figure 9: Demographics of Instagram users

sion, we recorded the content visible on their feeds, capturing logs of posts, suggested content, and advertisements. For this dataset, we restricted the collection to the 'Home' feed, although the tool is capable of recording data from other feeds, such as 'Reels' and 'Explore.'

The dataset includes demographic information such as political affiliation, ethnicity, gender, and age. Figure 9 provides an overview of the demographic distribution of the participants. Due to the small sample size, the dataset is skewed towards white, liberal, and middle-aged users.

On average, we recorded 106.7 items from each participant's feed (maximum: 685, median: 79). Approximately 17% of the content on average (median: 20%) were advertisements, while an average of 9% (median: 2.5%) consisted of suggested posts from accounts not followed by the participants. These figures highlight the extent to which algorithmically curated content, including ads and suggested posts, is present in users' feeds.

This dataset has significant potential for analysis, offering opportunities to explore questions such as the types of content different demographics are exposed to, the targeting patterns of advertisements, and the extent to which Instagram's feed resembles TikTok's algorithmic model. It also provides a novel approach to auditing and collecting data from Instagram. While this dataset may be limited in its current form, we hope it serves as a valuable starting point for the research community to develop further insights and methodologies.

the tool could be explored, this initial deployment focused on testing the app and evaluating the type of data that could be collected.

5 Discussion

The paper presents the development and deployment of data donation tools for various platforms, along with the datasets collected using these tools. Some of these tools and datasets are entirely novel, enabling access to previously unavailable data (e.g., WhatsApp), while others are experimental and hold potential for future research (e.g., Instagram). Still, others facilitate the collection of large-scale, user-specific data that was previously challenging to obtain or complement existing data collections (e.g., Telegram, YouTube).

These tools represent one approach to collecting social media data, not the definitive way. The hope is that they inspire researchers to rethink how social media data can be gathered and utilized using creative approaches like data donation. However, we must address a critical question: do we need yet another set of tools when many existing tools are no longer maintained? Inevitably, these tools may join the "graveyard" of social media research tools, becoming obsolete or unsupported. Yet, the concepts of data donation, user-specific data collection, and the datasets we release here remain valuable contributions to the research community.

A stark reality of working with social media platforms is that they ultimately hold the power. If platforms choose to obstruct these efforts, they have the resources to do so. While data donation is currently legal under existing terms of service, this may not always be the case. Even without legal challenges, practical issues, such as escalating costs and resource requirements, could hinder this approach.

Ethics have been central to this work, encompassing detailed informed consent processes, university ethics reviews, and comprehensive anonymization protocols. Critics concerned about the ethical implications of the data donation approaches we presented should consider the current alternative: relying on data purchased from third-party brokers, as seen in numerous studies published in prestigious journals such as Science and Nature (Eady et al. 2023; Guess, Nagler, and Tucker 2019; Hosseinmardi et al. 2021). These datasets often cost six figures, making them inaccessible to many researchers. Academic-led data donation offers a more ethical and equitable approach, addressing both cost and accessibility barriers.

Despite these efforts, practical challenges remain. Deploying these tools requires significant technical expertise, which is not universally available. This highlights the need for standardizing such tools to make them more accessible across disciplines, enabling researchers from diverse fields to deploy and adapt them for their studies. For a discussion on the costs associated with data donation, see Appendix Section A.2.

Finally, for data donation to be a sustainable and successful model, it is crucial to consider the value users receive beyond monetary compensation. Social media data donation, in particular, has the potential to educate users about their online behaviors and digital environments. While various

initiatives have explored this idea, a truly impactful application—a "killer" feature—has yet to emerge. Addressing this gap could unlock new possibilities for user engagement and the broader adoption of data donation practices.

6 Acknowledgments

I would like to thank the software engineers who helped create and maintain the tools: WhatsApp: Shreyash Jain, Ayushman Panda, Shlok Pandey, Facebook: Varun Matta, YouTube: Ankit Singh, Telegram: Paras Mehan, Harsh Mehta, Instagram: Kartik Ohri. This paper would not be possible with out the amazing contributions of the participants.

References

2024. Data Donation Infrastructure (D3I). https://datadonation.eu/. Accessed: 2025 Jan.

2024. Data Donation Lab. https://datadonation.uzh.ch/en/. Accessed: 2025 Jan.

Angus, D.; Obeid, A. K.; Burgess, J.; Parker, C.; Andrejevic, M.; Carah, N.; and Tan, X. Y. 2024. Enabling Online Advertising Transparency through Data Donation Methods. *Computational Communication Research*, 6(2): 1.

Araujo, T.; Ausloos, J.; van Atteveldt, W.; et al. 2022. OSD2F: An open-source data donation framework. *Computational Communication Research*, 4(2): 372–387.

Baumgartner, J.; Zannettou, S.; Squire, M.; and Blackburn, J. 2020. The pushshift telegram dataset. In *ICWSM*.

Boeschoten, L.; Ausloos, J.; Möller; et al. 2022. A framework for privacy preserving digital trace data collection through data donation. *Computational Communication Research*.

Boeschoten, L.; de Schipper, N. C.; et al. 2023. Port: A software tool for digital data donation. *Journal of Open Source Software*, 8(90): 5596.

Bond, S. 2021. NYU Researchers Were Studying Disinformation On Facebook. *The Company Cut Them Off. NPR (Aug. 2021)*.

Burroughs, B. 2014. Facebook and FarmVille: A digital ritual analysis of social gaming. *Games and Culture*.

Canevez, R. N.; Maikovska, K.; and Zwarun, L. 2024. Tactics and affordances in the mediatization of war: pro-Ukrainian cyber resistance on Telegram. *Digital War*, 5: 167–180.

Couto, J.; and Garimella, K. 2024. Examining (Political) Content Consumption on Facebook Through Data Donation. arXiv:2407.08171.

Eady, G.; Paskhalis, T.; Zilinsky, J.; Bonneau, R.; Nagler, J.; and Tucker, J. A. 2023. Exposure to the Russian Internet Research Agency foreign influence campaign on Twitter in the 2016 US election and its relationship to attitudes and voting behavior. *Nature communications*, 14(1): 62.

Feal, A.; Gleason, J.; Goel, P.; Radford, J.; Yang, K.-C.; Basl, J.; Meyer, M.; Choffnes, D.; Wilson, C.; and Lazer, D. 2024. Introduction to National Internet Observatory.

Fiesler, C.; Beard, N.; and Keegan, B. C. 2020. No robots, spiders, or scrapers: Legal and ethical regulation of data collection methods in social media terms of service. In *ICWSM*.

Garimella, K.; and Chauchard, S. 2024. WhatsApp Explorer: A Data Donation Tool To Facilitate Research on WhatsApp. arXiv:2404.01328.

Garimella, K.; Nayak, B.; Chauchard, S.; and Vashistha, A. 2024. Deciphering Viral Trends in WhatsApp: A Case Study From a Village in Rural India.

Gomez Ortega, A.; Bourgeois, J.; Hutiri, W. T.; and Kortuem, G. 2023. Beyond data transactions: a framework for meaningfully informed data donation. *AI & SOCIETY*.

Guess, A.; Nagler, J.; and Tucker, J. 2019. Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science advances*, 5(1): eaau4586.

Hase, V.; Jef, A.; Laura, B.; et al. 2024. Fulfilling data access obligations: How could (and should) platforms facilitate data donation studies? *Journal on Internet Regulation*.

Hosseinmardi, H.; Ghasemian, A.; Clauset, A.; Mobius, M.; Rothschild, D. M.; and Watts, D. J. 2021. Examining the consumption of radical content on YouTube. *Proceedings of the National Academy of Sciences*, 118(32): e2101967118.

Hummel, P.; Braun, M.; and Dabrock, P. 2019. Data donations as exercises of sovereignty. *The ethics of medical data donation*, 23–54.

Hutchinson, A. 2021. Instagram's Developing a TikTok-Style Vertical Feed for Stories.

Lazer, D.; Hargittai, E.; Freelon, D.; Gonzalez-Bailon, S.; Munger, K.; Ognyanova, K.; and Radford, J. 2021. Meaningful measures of human society in the twenty-first century. *Nature*, 595(7866): 189–196.

Lazer, D.; Santillana, M.; et al. 2020. The COVID States Project: A 50-state COVID-19 survey report# 26: Trajectory of COVID-19-related behaviors. *COVID States Project*.

Leibniz Institute for Media Research. 2024. Social Media Observatory. https://leibniz-hbi.de/en/. Accessed: 2025 Jan.

Lukito, J. 2024. Platform research ethics for academic research. *Center for media engagement.*, 30.

Norton, S.; and Shapiro, J. N. 2024. How to better study—and then improve—today's corrupted information environment — thebulletin.org.

Razi, A.; AlSoubai, A.; et al. 2022. Instagram data donation: a case study on collecting ecologically valid social media data for the purpose of adolescent online risk detection. In *CHI*.

Smuts, N. 2019. What Drives Cryptocurrency Prices?: An Investigation of Google Trends and Telegram Sentiment. In *HICSS*.

Stefana, A.; Dakanalis, A.; Mura, M.; Colmegna, F.; and Clerici, M. 2022. Instagram Use and Mental Well-Being: The Mediating Role of Social Comparison. *Journal of Nervous and Mental Disease*.

Stocking, Galen and van Kessel, and others. 2020. Many Americans Get News on YouTube, Where News Organizations and Independent Producers Thrive Side by Side.

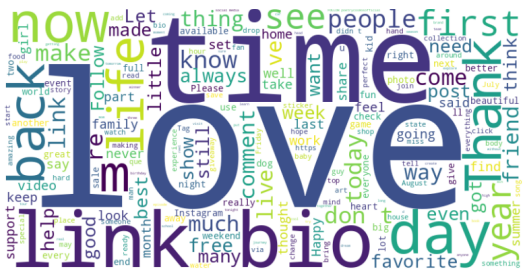

Figure 10: Word cloud of posts from Instagram

ThinkImpact. 2024. Telegram Statistics 2024. Accessed: 2025-01-14.

van Driel, I. I.; Giachanou; et al. 2022. Promises and pitfalls of social media data donations. *Communication Methods and Measures*, 16(4): 266–282.

Walker, M.; and Eva Matsa, K. 2021. News consumption across social media in 2021. *Pew Research Center*.

Walther, S.; and McCoy, A. 2021. US Extremism on Telegram: Fueling Disinformation, Conspiracy Theories, and Accelerationism. *Perspectives on Terrorism*, 15(2).

A Appendix

A.1 Word clouds of Instagram posts

Figure 10 shows the word cloud of the captions of posts.

A.2 Costs of data donation

The primary cost in data donation lies in the development and maintenance of the tools. While deploying the tools and recruiting participants is relatively inexpensive, building and sustaining these tools over time presents a significant challenge. For example, the Facebook data donation initiative described in this paper incurred a participant recruitment cost of approximately $2,500 USD. Given the ability to compute nationally representative statistics about Facebook usage from this dataset, this is a relatively low cost. Similarly, the data collection for Telegram and YouTube cost under $2,000 each. Participants were compensated at an average rate of $30 per hour for their time, making this approach relatively cost-effective compared to traditional methods like purchasing data from third-party brokers.

Even though the main cost is the initial development, the real challenge is not in the initial development but in ensuring the long-term viability of such tools. Development requires substantial technical expertise and financial resources, and ongoing maintenance often becomes infeasible as platforms evolve or impose restrictions. For instance, the Facebook tool described here is currently nonfunctional due to changes on the platform, and while repairing it might be possible, it is not practical within existing constraints. Such adversarial dynamics—where tools operate against the interests of large, resourceful companies—make sustained maintenance a daunting task for individual researchers.

To address these issues, it is imperative that the academic community comes together to standardize, build, and maintain these tools collaboratively. Organizations like the National Internet Observatory could play a pivotal role in spearheading such efforts. However, priorities and expertise often vary; for instance, my focus on mobile platforms may differ from others' interests in desktop solutions or specific datasets. Despite these differences, creating a unified framework for tool development and maintenance is essential for the long-term success of data donation initiatives.

Institutions like the Inter-university Consortium for Political and Social Research (ICPSR) or Princeton's Social Media Accelerator (Norton and Shapiro 2024) could also take on the responsibility of maintaining these tools. These organizations already have experience managing large-scale data infrastructure and could provide the technical, legal, and financial support necessary to sustain data donation efforts. By forming dedicated teams to ensure these tools remain functional and accessible, such institutions could significantly advance the research community's ability to study social media platforms.