Article

Reflecting on the State of Digital Media Data Research

February 6, 2023

In January 2023, Georgia Governor Brian Kemp and Texas Governor Greg Abbott banned TikTok on state-owned devices, effectively halting university research on TikTok. And in February, Twitter announced it would end the free-tier of its public APIs, removing API access for many researchers. These actions are the latest in a string of challenges for researchers studying digital media data, which includes (but is not limited to) social media content, digital trace data, and digitized data.

And they're not likely to be the last. As the conditions for studying digital media data change rapidly, researchers must carefully consider their work's practical, ethical, and legal challenges. These challenges are not limited to one part of digital media data research: at every stage of a study (collection, analysis, and reporting), reflexivity is necessary.

In an effort to capture the trends in digital media data research, the Media & Democracy Data Cooperative is releasing a report, The State of Digital Media Data Research, 2023, highlighting the key challenges that digital media data (DMD) researchers face. The report is also a call to action for greater collaboration, transparency, preparation, and consistency within the interdisciplinary and growing DMD research community. This report synthesizes the opinions and experiences of the data cooperative, as well as insights from the 2022 Digital Data Conference.

Identifying Challenges

At the data collection stage, researchers have a variety of strategies to collect digital media data, from official APIs to data scraping tools. While expansive, researchers must be aware that different collection approaches bring different legal and ethical boundaries. Official APIs, for example, can make the research process easier. However, without policies like the DSA, these resources rely on the waning generosity of platforms. More unofficial data collection tactics (e.g. web scraping) create different legal and ethical issues for the researcher, including copyright and terms of service violations. Another concern with data collection is the over-reliance on a few platforms to collect publicly available data, a

problem that has only grown with Twitter's recent announcements about API access. A third challenge is the veracity of collected data, as any collection may have unexpected gaps in what is collected, by virtue of rate limit errors, scraping issues, and deleted content.

Furthermore, the ability to collect and store this data is not available to all researchers. While some tools are more widely accessible, particularly for those collecting "small data," larger data collections and multi-modal data may require costly data storage. As a result, there is a disparity between larger, more well-resourced teams and individual researchers with limited research resources, affecting what research questions are asked and answered.

These challenges also extend to analyzing digital media data, including issues with the cost of analysis, decontextualized analysis, and traumatic digital content, highlighting the need to financially and institutionally support digital media data research. Similar to concerns about data collection, the analysis of large digital media datasets can be prohibitively expensive, from charges for out-of-the-box analytics tools to the fees associated with computing resources. To handle data of this size and scale, researchers have relied on computational strategies. However, this can decontextualize the data (sometimes intentionally so), masking nuances of naturally-occurring human behavior and communication. This tension highlights the continued need for qualitative research that situates digital media in our everyday lives. And thirdly, scholars are often interested in studying unwanted content on social media. This content includes (but is not limited to) hate speech, incivility, mis- and disinformation, and Child Sexual Abuse Material. This can be challenging for both data collection–since data are often deleted–and data analysis, as researchers studying this content may encounter secondary traumatic stress. The concern of traumatic content is particularly acute among early career scholars, graduate students, and undergraduate research assistants who often label such data.

However, the research does not stop at the analysis level. Increasingly, researchers are motivated, both personally and professionally, to share their data in some capacity. But these efforts are easier said than done. Researchers must balance data practices across multiple dimensions: a desire for open access to the data, replicability, and users' right to privacy and informed consent. Platform restrictions to sharing data exacerbate these concerns: researchers must balance open science principles with the issue of user privacy and limitations of Terms of Service and Data Use Agreements from the platforms that grant access to digital media data.

When broadly considering the digital media data research process (collection, analysis, and sharing), it is important to note that these challenges are not discrete–they inform one another. How one researcher collects and shares data may impact a future researcher who uses a shared version of the data that analyzes or aggregates content to preserve user

privacy or meet the expectations of Terms of Service agreements.

Recommendations for the Future

We highlight four core principles that can help advance the study of digital media data in ever-changing times.

First, we encourage researchers to collaborate. A plurality of approaches is needed to conduct digital media data research, including independent and collaborative modes of data collection and large-scale computational/quantitative and rich qualitative methods of analysis.

Second, researchers must be transparent about their work. While there may be restrictions to the data being fully published, researchers can still be transparent about their collection, analysis, and sharing approach, whether making their code and research materials accessible and reproducible or providing cost estimates regarding their research.

Third, in the face of an ever-changing digital media ecology, researchers must also be prepared. The data we study and our approaches to studying them must continually be adapted to respond to these changes. For this reason, scholars must build alternative data collection and risk assessment plans into their study's design.

And finally, researchers should be consistent in their approach to assessing ethical risks and challenges to their study. Understanding that no one-size-fits-all approach will work for digital media data research, an ethical framework remains necessary to give new researchers a sense of risk considerations of their work. To that end, researchers should develop cross-disciplinary norms for how we conduct ethical risk assessments for digital media data research and how to ethically anonymize and share that data.

Conclusion: Ensuring a Bright Future for Digital Media Data Research

Research on digital media data spans many disciplines and methodologies. Despite differences in our theoretical and empirical approaches, researchers must be united in their efforts to combat the ever-evolving challenges of digital media data research. As the discipline grows, collaboration, transparency, preparation, and consistency are necessary to ensure that researchers work towards a more ethical, equitable, and rigorous standard for digital media data research.