Instagram Data Donation: A Case for Partnering with Social Media Platforms to Protect Adolescents Online

Xavier Caddle
University of Central Florida
Orlando, U.S.A
xavier.caddle@knights.ucf.edu

Ashwaq AlSoubai
University of Central Florida
Orlando, Florida, U.S.A
atalsoubai@knights.ucf.edu

Afsaneh Razi
University of Central Florida
Orlando, Florida, U.S.A
afsaneh.razi@knights.ucf.edu

Seunghyun Kim
Georgia Institute of Technology
Atlanta, Georgia, U.S.A
seunghyun.kim@gatech.edu

Shiza Ali
Boston University
Boston, Massachusetts, U.S.A
shiza@bu.edu

Gianluca Stringhini
Boston University
Boston, Massachusetts, U.S.A
gian@bu.edu

Munmun De Choudhury
Georgia Institute of Technology
Atlanta, Georgia, U.S.A
munmund@gatech.edu

Pamela Wisniewski
University of Central Florida
Orlando, Florida, U.S.A
pamwis@ucf.edu

ABSTRACT

Social Media platforms collect data about their users, including basic demographic information (name, age, gender) as well as their interactions with other users. This rich multi-modal data could potentially be leveraged to identify trends that can be used to support the safety and well-being of adolescents. However, researchers who try to access such data are faced with a number of challenges getting access to such social media data. In this paper we present some the challenges we faced when creating a data set for our NSF funded project to improve adolescents' online safety by developing human-centered algorithms for online risk detection, which led us to our position: Social media platforms should work alongside researchers to promote ethical and responsible research that benefits youth and works to prevent the risks they encounter online. Our goal in attending this workshop is to co-create best practices for facilitating ways for academics to work with social media companies to reduce the barriers posed when using social media as a research site.

ACM Reference Format:
Xavier Caddle, Ashwaq AlSoubai, Afsaneh Razi, Seunghyun Kim, Shiza Ali, Gianluca Stringhini, Munmun De Choudhury, and Pamela Wisniewski. 2021. Instagram Data Donation: A Case for Partnering with Social Media Platforms to Protect Adolescents Online. In *Proceedings of Social Media as a Design and Research Site in HCI: Mapping Out Opportunities and Envisioning Future Uses (CHI 2021).* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

In an effort to capitalize on the wealth of data collected by social media platforms, researchers often leverage this data in ways that strive to benefit society [2–4, 18]. These efforts are bolstered when social media platforms provide fair and transparent access to this data. An exemplar of open access for social media research is Twitter, which recently revamped their Application Programming Interface (API), adding a new access level for academic researchers [5]. Another example is Facebook's CrowdTangle tool which uses public content from Instagram and Facebook Pages to provide users with some insights into the system [20]. However, most social media platforms do not provide means for good actors, such as researchers, to access data that could be used to help protect individuals who use their platforms. Our National Science Foundation (NSF) funded Partnerships for Innovation (PFI) project is a case-in-point for demonstrating the challenges researchers face when trying to leverage social media data for good.

This hesitancy for social media companies to engage with researchers is not without good reason. After the Cambridge Analytica scandal [13], the public has become increasingly wary of what data social companies collect and how that data is used. After learning of being surveilled, Americans have changed the way in which they interact online [16]. One study found that 50% of Americans do not trust social media with protecting their data [15]. The European Union (EU) has also created legislation (i.e., GDPR) to outline how data belonging to their citizens should be handled impose penalties for non-compliance [1]. The legislation affords EU citizens with certain rights including the right of access to their data and the right to restrict the processing of their data [23]. Competing interests may be another reason that prevents social media platforms from working more closely with researchers. From a business perspective, there exists tension between the need to drive profits, mitigate liability, and the need to be a responsible actor in society [22]. Since researchers must conduct their research without conflicts of interest, their research findings cannot be beholden to the interests of social media corporations. This causes conflict when

CHI 2021, May 15–16, 2021, Online Virtual Conference)

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

engaging in social media research as our findings could negatively impact the interests of the company; therefore, making it in the best interest of the social media company to disengage and protect their own interests. In summary, the tensions between open access data for the purpose of research and these concerns have created significant barriers for using social media as a research site, particularly when private (e.g., direct messages), rather than public (e.g., public posts), social media data is needed to advance important and timely research agendas.

2 USING SOCIAL MEDIA AS A RESEARCH SITE FOR ADOLESCENT ONLINE SAFETY

Teen internet access and social media use has jumped significantly in the United States and worldwide. In 2012, Common Sense Media found that 4 out of 10 US teens had smartphones [19]. This number increased to 9 out 10 teens in 2018, meaning that approximately 90 percent of US teens have access to the internet from their own phone. The same report noted that that in 2018, 81% of teens used social media and 70% of those teens admitting to doing so at least once a day. As teens use social media, they often share private information about themselves and interact with a wide array of people that makes them susceptible to online risks. With increased internet usage, comes the potential increase of exposure to pornography, sexual solicitation, cyberbullying, mental-health issues, suicide, and other types of online risks [11]. One study reported that 64% of teens 13-17 in age were exposed to hate speech online and also found that exposure to racist content online increased by 9% between the years 2012 and 2018 [19].

Given these problems, the goal of our project is to use a human-centered approach to develop and open source, risk detection algorithms for adolescent online risk detection to protect youth online. Our current study aims to create more accurate teen online risk detection models by using both public and private social media data. We believe that by utilizing both the public and private conversations, we can create a more accurate models of teens' online interactions. This in turn will allow us to better identify when teens are exposed to risks. To achieve this goal, teens submit their data file to us and are asked to flag which messages represent risky encounters. This process alleviates the feeling of privacy intrusion since the data is downloaded and submitted by the teen. The teen explicitly consents to sharing their data for use in our research.

In creating the dataset for our study, we were faced with many challenges which we will present in this paper. These challenges lead us to our position: We believe social media platforms should work more closely with researchers to facilitate high-quality studies that are ethically designed. We put forward this stance due to our experience in our current study which aims to use public and private social media content.

3 CHALLENGES AND BARRIERS

While conducting our research we were faced with challenges ranging from accessing the social media data, obtaining consent to use the social media data, to some ethical issues regarding the sharing of certain media files. In our study we have found that teens prefer to pick what data is shared with other parties when the data is of a private nature [2, 17]. The following sections elaborate on these challenges.

3.1 Ethical Challenges

Collecting data from social media is a sensitive topic; however, when the data is collected from teens, additional considerations are required to use the collected data in an ethical way. Participants' privacy is a high priority when dealing with teens both during the study after when disseminating the results. In our project, after obtaining informed consent (parental consent for teens under 18 and teens' assent), we ask teens and young adults (13-21 year old) to voluntarily request and download their data file from Instagram and consent to us using the data to improve machine learning models to detect online risks. The Human Computer Interaction (HCI) community identified adolescents (ages 13-17) as an "understudied and vulnerable population" [14]. The HCI community has also defined multiple challenges that researchers can face when dealing with teens such as how the teens' age can make them more vulnerable by law which can also create a power imbalance between the researchers and teens. Another identified challenge is making sure that teens understand the information for consent, "maintaining confidentiality of teens' data, and protecting youth from harm or abuse" [6, 9, 10]. In research we have recently presented, we found the need to ensure the privacy and confidentiality of teens' social media data [2]. We found that this even extends to revealing the identity of the teens' message recipient since they feared getting them in trouble [4]. Teens' and their parents place great importance on the need to protect their identity. On average, teens positively responsive to sharing their social media data if they have full control over the data itself, want to share their personal online experiences with others, and would like to contribute to society to help others in this area [2, 4]. The collected data in our project is considered "restrictive data"; therefore, we have a very protective set of instructions for all members of this project (including: programmers, data verifiers, and data annotators) to keep teens' data safe. The instructions covered all technical means such as keeping the dataset in a virtual private cloud, encrypting the data transmission from users' browsers to the server, restricting access to the server and dataset to only the authorized people, and not using any cloud-based APIs for data analysis.

3.2 Legal Challenges

In our project, we have an additional complexity layer since we are working on detecting online risk behaviors (e.g., harassment, sexual solicitations, etc.) to protect teens. This includes collecting very sensitive interactions that have the potential of putting minors at legal risk. Asking teens to share their online sexual risk experiences with researchers, for instance, doubled the responsibility for researchers [4]. For example, it is against federal law to have teens share nudes even if it is for research purposes; therefore, the Institutional Review Board (IRB) approval made it clear in our project to ask teens to remove nude images before sending it to us. This is compounded by the fact that researchers are legally bound to report any child abuse or imminent risks detected during our studies. This is also troublesome for the social media corporations as disclosing these incidents necessitates disclosing which social media platform where the transgressions occur.

Another challenge is that social media corporations seem to take steps to block the scraping of data from their platform as there are request limits for API requests from a single internet address. While the legality of scraping has been contested, due to lack of researcher access, researchers have had to resort to some process similar to scraping in order to collect social media data. Downloading media from incomplete data files is one such process. In practice we have also found that social media corporations also block API requests from known cloud services providers, thereby rendering the data collection process arduous.

3.3 Technical Challenges

Although fast-pace updates are common in industry settings, having constant changes brings challenges to researchers as well as participants. During our project, Instagram changed the formatting of the data file more than three times, which was a roadblock for us since we had to change our back-end code for processing these files. In addition, these format changes are also confusing for participants to find the information that they need to view on their data file. Even though Instagram provides the data file to users, we have noted challenges in that some private media (images/videos) are not directly available in the Instagram data file but are replaced with links to the actual post.

Accessing these links requires logging into the platform associated with the original post; in cases where the user, the teen who shared the post in this case, only has access to the media, it is difficult for the researchers to obtain the original image or video from the shared post. This is further solidified in the Terms of Service (ToS) set by social media platforms. For instance, Facebook specifically states that API users can only use data retrieved from its “oEmbded” endpoint for display purposes only [8].

4 POTENTIAL PATHS FORWARD

Social media platforms can help mitigate the ethical and legal challenges for researchers while maintaining teens' privacy and confidentiality. We previously mentioned Twitter [5] and CrowdTangle [20]. Facebook has also started a program called "Data for Good"[1] which includes tools built from privacy-protective data on Facebook to assist researchers. If social media platforms partner with researchers, then we can jointly develop best practices for using social media as a research site.

Social media platforms can also help researchers by facilitating data access more seamlessly. For example, social media platforms could collaborate with researchers to securely share data as long as proper parental consent/teen assent is obtained. They could also provide opt-in consent processes to the end users, teens and their parents in this case, that explains in detail the data content that will be collected as well as how the collected data will be used. Providing complete data files to users on request can also help with the transparency of what will or can be actually shared to researchers. Additionally, communicating with researchers when data format changes occur and why would be immensely helpful. Such approaches would further strengthen the potential of social media as a data source for research on adolescent online risk detection as well as provide researchers with an effective and ethical way of collecting data.

5 BENEFITS OF WORKSHOP ATTENDANCE

The Ph.D. students supported by the NSF grant could greatly benefit from attending this workshop to learn from senior researchers how to ethically and meaningfully engage with social media as a research site. We have encountered numerous challenges and would like to share these challenges and co-create solutions with other researchers who have encountered similar obstacles. By working together as a community of researchers, we can forge a path that makes it more feasible to use social media as a research site.

6 CONCLUSION

We have presented some of the challenges faced by researchers seeking to use social media data in their research efforts. We present the position that social media companies should be more willing to work with researchers to leverage their data for research that promotes the online safety and well-being of youth. We understand that our position is not without challenges or controversy. Similar to how social media companies pushed the idea of “frictionless sharing” [12] that eroded privacy norms for decades, we understand why social media companies may be hesitant to engage with researchers who examine youth risk behavior on their platforms. However, social media companies also have a social responsibility to protect vulnerable users. For instance, the U.S. recently passed legislation that extended the liability for human trafficking to social media [21]. Now that social media companies are embracing more privacy-focused agendas [24] and even funding privacy-related research [7], these companies also need to start embracing research that promotes the online safety and well-being of users, particularly minors.

ACKNOWLEDGMENTS

This research is partially supported by the U.S. National Science Foundation under grant IIP-1827700. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the research sponsors.

REFERENCES

[1] [n.d.]. GDPR and Social Media: What Data Protection and Privacy Mean for Social Media Marketers. https://influencermarketinghub.com/gdpr-social-media/. (Accessed on 02/16/2021).

[2] Zainab Agha, Neeraj Chatlani, Afsaneh Razi, and Pamela Wisniewski. 2020. Towards Conducting Responsible Research with Teens and Parents regarding Online Risks. *CHI EA '20: Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (April 2020), 1–8.

[3] AMII. 2020. *Tracking Mental Health During the Coronavirus Pandemic.* Retrieved February 11, 2021 from https://www.amii.ca/latest-from-amii/mental-health-coronavirus/

[4] Karla Badillo-Urquiola, Zachary Shea, Zainab Agha, Irina Lediaeva, and Pamela Wisniewski. 2021. Conducting Risky Research with Teens: Co-designing for the Ethical Treatment and Protection of Adolescents. *Proceedings of the ACM on Human-Computer Interaction* (Jan. 2021). https://doi.org/10.1145/3432930

[5] Ian Cairns. 2020. *Introducing a new and improved Twitter API.* Retrieved February 10, 2021 from https://blog.twitter.com/developer/en_us/topics/tools/2020/introducing_new_twitter_api.html

[1] For more information on this tool see here: https://research.fb.com/data/

[6] Jóhanna Einarsdóttir. 2007. Research with children: Methodological and ethical challenges. *European early childhood education research journal* 15, 2 (2007), 197–211.

[7] Facebook. 2020. *People's Expectations and Experiences with Digital Privacy request for proposals.* Retrieved February 21, 2021 from https://research.fb.com/programs/research-awards/proposals/peoples-expectations-and-experiences-with-digital-privacy-request-for-proposals/

[8] Facebook. 2021. *oEmbed - Instagram Platform.* Retrieved February 11, 2021 from https://developers.facebook.com/docs/instagram/oembed

[9] Rosie Flewitt*. 2005. Conducting research with young children: Some ethical considerations. *Early child development and care* 175, 6 (2005), 553–565.

[10] Anne Graham, Mary Ann Powell, and Nicola Taylor. 2015. Ethical research involving children: Encouraging reflexive engagement in research with children and young people. *Children & Society* 29, 5 (2015), 331–343.

[11] Sonia Livingstone and Leslie Haddon. 2008. Risky experiences for children online: charting European research on children and the Internet. *Children and Society* 22 (July 2008), 314–323. http://www.wiley.com/bw/journal.asp?ref=0951-0605

[12] Harry McCracken. 2011. *Did Facebook Just Change Social Networking Forever?* Retrieved February 21, 2021 from http://content.time.com/time/business/article/0,8599,2095516,00.html

[13] Sam Meredith. 2018. *Here's everything you need to know about the Cambridge Analytica scandal.* Retrieved February 19, 2021 from https://www.cnbc.com/2018/03/21/facebook-cambridge-analytica-scandal-everything-you-need-to-know.html

[14] Erika S Poole and Tamara Peyton. 2013. Interaction design research with adolescents: methodological challenges and best practices. In *Proceedings of the 12th International Conference on Interaction Design and Children.* 211–217.

[15] LEE RAINIE. [n.d.]. Americans' complicated feelings about social media in an era of privacy concerns. https://www.pewresearch.org/fact-tank/2018/03/27/americans-complicated-feelings-about-social-media-in-an-era-of-privacy-concerns/. (Accessed on 02/16/2021).

[16] LEE RAINIE and MARY MADDEN. [n.d.]. Americans' Privacy Strategies Post-Snowden. https://www.pewresearch.org/internet/2015/03/16/americans-privacy-strategies-post-snowden/. (Accessed on 02/16/2021).

[17] Afsaneh Razi, Zainab Agha, Neeraj Chatlani, and Pamela Wisniewski. 2020. Privacy Challenges for Adolescents as a Vulnerable Population. *Networked Privacy Workshop of the 2020 CHI Conference on Human Factors in Computing Systems* (April 2020). https://doi.org/10.2139/ssrn.3587558

[18] Andrew G. Reece and Christopher M Danforth. 2017. Instagram photos reveal predictive markers of depression. *EPJ Data Science* 6, Article 15 (Aug. 2017). https://doi.org/10.1140/epjds/s13688-017-0110-z

[19] Victoria Rideout and Michael B Robb. 2018. *Social Media, Social Life. Teens Reveal Their Experiences.* Retrieved February 10, 2021 from https://www.commonsensemedia.org/sites/default/files/uploads/research/2018_cs_socialmediasociallife_fullreport-final-release_2_lowres.pdf

[20] Naomi Shiffman. 2021. *Social Media, Social Life. Teens Reveal Their Experiences.* Retrieved February 18, 2021 from https://help.crowdtangle.com/en/articles/4558716-understanding-and-citing-crowdtangle-data

[21] William M. Sullivan Jr. and Leonardi Fabio. 2018. *Bill Expands Corporate Liability for Human Trafficking to Social Media Companies.* Retrieved February 21, 2021 from https://www.pillsburylaw.com/en/news-and-insights/bill-expands-corporate-liability-for-human-trafficking-to-social-media-companies.html

[22] Taylor Tepper. 2018. *Milton Friedman On The Social Responsibility of Business, 50 Years Later.* Retrieved February 18, 2021 from https://www.forbes.com/advisor/investing/milton-friedman-social-responsibility-of-business/

[23] Ben Wolford. [n.d.]. What is GDPR, the EU's new data protection law? https://gdpr.eu/what-is-gdpr/. (Accessed on 02/16/2021).

[24] Mark Zuckerberg. 2019. *A Privacy-Focused Vision for Social Networking.* Retrieved February 21, 2021 from https://www.facebook.com/notes/2420600258234172/?comment_id=923865591403095