Instagram Data Donation: A Case Study on Collecting Ecologically Valid Social Media Data for the Purpose of Adolescent Online Risk Detection

AFSANEH RAZI, University of Central Florida, U.S.A
ASHWAQ ALSOUBAI, University of Central Florida, U.S.A
SEUNGHYUN KIM, Georgia Institute of Technology, U.S.A
NURUN NAHER, University of Central Florida, U.S.A
SHIZA ALI, Boston University, U.S.A
GIANLUCA STRINGHINI, Boston University, U.S.A
MUNMUN DE CHOUDHURY, Georgia Institute of Technology, U.S.A
PAMELA J. WISNIEWSKI, University of Central Florida, U.S.A

In this work, we present a case study on an Instagram Data Donation (IGDD) project, which is a user study and web-based platform for youth (ages 13-21) to donate and annotate their Instagram data with the goal of improving adolescent online safety. We employed human-centered design principles to create an ecologically valid dataset that will be utilized to provide insights from teens' private social media interactions and train machine learning models to detect online risks. Our work provides practical insights and implications for Human-Computer Interaction (HCI) researchers that collect and study social media data to address sensitive problems relating to societal good.

CCS Concepts: • **Human-centered computing → Human computer interaction (HCI)**; • **Empirical studies in HCI**;

Additional Key Words and Phrases: Adolescents, Teens, Datasets, Instagram, Data Collection

ACM Reference Format:
Afsaneh Razi, Ashwaq AlSoubai, Seunghyun Kim, Nurun Naher, Shiza Ali, Gianluca Stringhini, Munmun De Choudhury, and Pamela J. Wisniewski. 2021. Instagram Data Donation: A Case Study on Collecting Ecologically Valid Social Media Data for the Purpose of Adolescent Online Risk Detection. 1, 1 (October 2021), 12 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Youth are avid technology and social media users. According to Pew Research [3], 45% of teens in the U.S. are constantly connected to the internet. Meanwhile, 72% of these youth are Instagram users. While using social media platforms provide benefits, such as social connections, learning, and creativity [32], these platforms also expose them to online

Authors' addresses: Afsaneh Razi, afsaneh.razi@knights.ucf.edu, University of Central Florida, 4000, Orlando, Florida, U.S.A, 32816; Ashwaq AlSoubai, University of Central Florida, 4000, Orlando, Florida, U.S.A, atalsoubai@Knights.ucf.edu; Seunghyun Kim, Georgia Institute of Technology, 30318, Atlanta, Georgia, U.S.A, seunghyun.kim@gatech.edu; Nurun Naher, University of Central Florida, 4000, Orlando, Florida, U.S.A, nurun.naher@Knights.ucf.edu; Shiza Ali, Boston University, 02215, Boston, Massachusetts, U.S.A, shiza@bu.edu; Gianluca Stringhini, Boston University, 02215, Boston, Massachusetts, U.S.A, gian@bu.edu; Munmun De Choudhury, Georgia Institute of Technology, 30318, Atlanta, Georgia, U.S.A, munmund@gatech.edu; Pamela J. Wisniewski, University of Central Florida, 4000, Orlando, Florida, U.S.A, pamwis@ucf.edu;

risks, such as cyberbullying [7, 34], sexual risks, and exposure to inappropriate content [5, 20, 30]. A growing body of literature has studied the negative effects of social media on youth, including mental-health, self-harm, and suicide ideation [8, 27]. However, the adolescent online safety literature suffers from reliance on self-reported data from surveys/interviews [29], which are prone to subjective assessments, recall, and hindsight biases [2].

In recent years there has been considerable interest e.g., [18, 22, 35] in detecting and/or mitigating [38] these online risks to keep youth safe online. To make online risk detection systems timely, scalable, and most importantly, accurate, it is crucial that the detection models are built upon ecologically valid datasets that depict the target users (i.e., youth) [9]. However, the majority of automated approaches for online risk detection on social media are based on datasets that do not accurately represent young social media users [23, 31]. A systematic review of the past literature on sexual risk detection revealed how studies have been skewed towards public datasets, which digress from the private discourse of online communication, where most sexual risks incidents occur [31]. Razi et al.'s review further highlighted how past studies on sexual risk detection were based primarily on a single dataset comprised of conversations between predators and adults posing as children, which fell short of representing the real victims of sexual predation. Similarly, Kim et al.'s literature review on cyberbullying detection emphasized how ground truth should be determined through direct involvement with stakeholders (i.e., youth victims of cyberbullying) [23]. Incorporating the perspectives of victims is crucial, as it enables the machine learning models to catch implicit inferences to the said risk [24]. The heavy reliance on external annotators with lack of first-person perspective when establishing the ground truth for training the detection models have been criticized by the aforementioned reviews [23, 31], which advocated for a more human-centered approach to strengthen the validity of these datasets. Previous study on the methodological gaps in predicting mental states has also emphasized how using proxy signals without any self-reported labels could lead to critical misclassifications and deprive the credibility of the predictions [14].

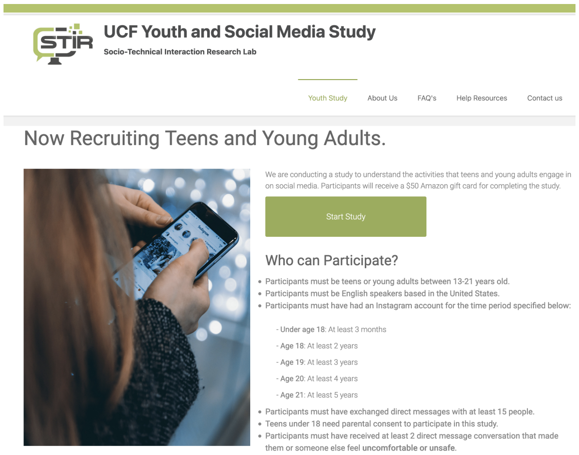
Establishing ecologically valid datasets, as well as considering different perspectives of the key stakeholders, when constructing ground truth fall under the approach of human-centered machine learning (HCML). HCML emphasizes that machine learning incorporates human-centered design and transparency for the sake of explaining usages in real-life scenarios as well as any potential to cause harm [10, 17]. Such practices to provide meaning and interpretability to the data-driven decisions are important as they provide a deeper understanding on the impacts of the machine learning models on the humans. As part of an National Science Foundation (NSF) funded Partnerships for Innovation (PFI) program, we built an online system to collect youth social media data integrated with their self-reported data.Our research project makes dataset and artifact contributions [40] to the fields of Human-Computer Interaction (HCI), adolescent online safety, Human-centered Machine Learning (HCML), and Social Computing (SC).Our work utilizes human-centered design to build an ecologically valid dataset based on digital trace data from youth and their perspective of online risks. We do this by asking youth (ages 13-21) to donate their personal Instagram data, including their private messages, for the purpose of research. Then, we have these youth participants annotate their own private messages for situations that made them or someone else feel uncomfortable or unsafe. In addition to collecting social media trace data, we also collected self-reported pre-validated survey constructs to assess our participants social media usage, online risk experiences, mental health status, and demographic information. Finally, we took great care to design this study in a way that protected the privacy of our participants. In this paper, we explain our design and study decisions, lessons learned through the design, development, and the data collection process. In addition, our findings provide implications for future data collection and research.

2 STUDY DESIGN AND DATA COLLECTION

We collected pre-validated survey measures and real world social media data from youth. We aimed to create a robust training dataset using the youths' social media data and establish ground truth labels for risks by utilizing participants' perspectives. We designed and developed a secure web-based system, where participants could fill out an online survey about their social media use, personal and online risk experience, download their Instagram data file and upload it in our system, and flagged their private message conversations that made them feel uncomfortable or unsafe. We selected Instagram as the platform for data collection as it was popular among youth (72% of teens use Instagram) [3]. Instagram and YouTube are the top social media platforms being used by half of U.S. teens ages 13 to 17 [3]. Instagram provides a way for users to download their data, as General Data Protection Regulation (GDPR) [16] mandates social media companies to provide options for users to download their personal data.

2.1 Data Collection Design and Approach

Figure 1 displays the main page of the website including the eligibility criteria. Through a Qualtrics survey, we recruited participants of age 13-21 who were: 1) English speakers based in the United States, 2) Had an active Instagram account currently and for at least 3 months during the time they were a teen (ages 13-17), 3) Exchanged DMs with at least 15 people, and 4) Had at least 2 DMs that made them or someone else feel uncomfortable or unsafe.


STIR
UCF Youth and Social Media Study
Socio-Technical Interaction Research Lab
Youth Study
About Us
FAQ's
Help Resources
Contact us
Start Study

Now Recruiting Teens and Young Adults.

We are conducting a study to understand the activities that teens and young adults engage in on social media. Participants will receive a $50 Amazon gift card for completing the study.

Who can Participate?

- Participants must be teens or young adults between 13-21 years old.
- Participants must be English speakers based in the United States.
- Participants must have had an Instagram account for the time period specified below:

  - **Under age 18**: At least 3 months
  - **Age 18**: At least 2 years
  - **Age 19**: At least 3 years
  - **Age 20**: At least 4 years
  - **Age 21**: At least 5 years

- Participants must have exchanged direct messages with at least 15 people.
- Teens under 18 need parental consent to participate in this study.
- Participants must have received at least 2 direct message conversation that made them or someone else feel **uncomfortable or unsafe**.

Fig. 1. Instagram Data Donation Main Page

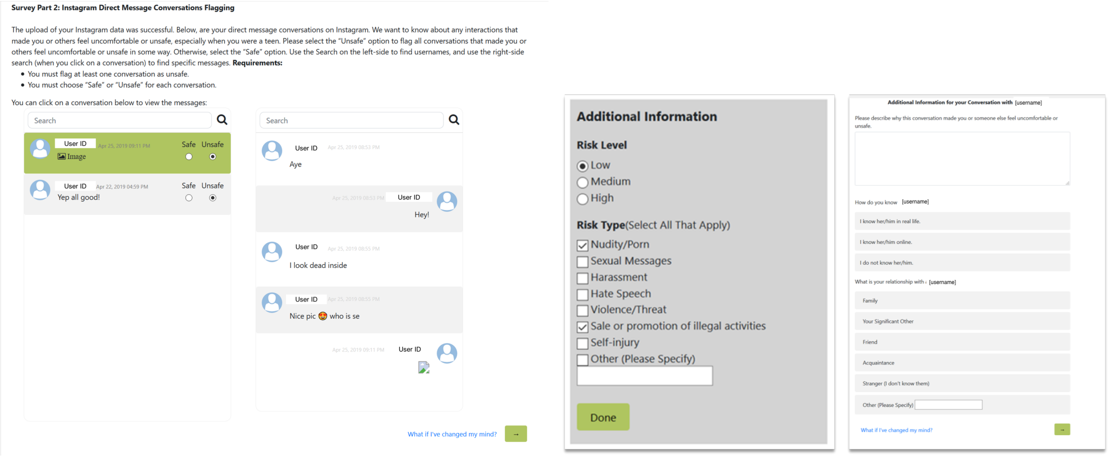
2.2 Consent and Assent

We carefully designed the study to only send parental consent and teen assent after participants passed the eligibility requirements. Following approval from the Institutional Review Board (IRB) of the authors' institutions, participants under the age of 18 were required to obtain parental consent prior to participating in the study. To make sure that teens are willingly participating in our study, we also included teen assent forms for those under 18. If they were older than 18, they were required to fill out the adult consent form. In the consent and assent forms, we included information about the research, research process, potential benefits and risks for participating in this research. Additionally, we also clarified what information would be collected and how it will be stored and protected, and anything else that participants needed to know to participate in our study.

2.3 Survey Measures

In this study, we aimed to understand different dimensions associated with social media experiences and online risks such as sexual risks, mental health issues, and cyberbullying. We gathered pre-validated survey measures to understand these risk behaviors. The main goal is also to associate this survey data with their Instagram data to understand their real world online social media interactions better.

2.3.1 Social Media Use. We asked participants about their social media usage to understand how they spend time on Instagram and other social media. These questions include measures from Facebook Intensity Scale by Ellison et al. [13].



Survey Part 2: Instagram Direct Message Conversations Flagging

The upload of your Instagram data was successful. Below, are your direct message conversations on Instagram. We want to know about any interactions that made you or others feel uncomfortable or unsafe, especially when you were a teen. Please select the "Unsafe" option to flag all conversations that made you or others feel uncomfortable or unsafe in some way. Otherwise, select the "Safe" option. Use the Search on the left-side to find usernames, and use the right-side search (when you click on a conversation) to find specific messages. **Requirements:**

- You must flag at least one conversation as unsafe.
- You must choose "Safe" or "Unsafe" for each conversation.

You can click on a conversation below to view the messages:

User ID Apr 25, 2019 09:11 PM
Image
Safe ○ Unsafe ●

User ID Apr 22, 2019 04:59 PM
Yep all good!
Safe ○ Unsafe ●

User ID Apr 25, 2019 08:53 PM
Aye

Apr 25, 2019 08:53 PM User ID
Hey!

User ID Apr 25, 2019 08:55 PM
I look dead inside

User ID Apr 25, 2019 08:55 PM
Nice pic 😍 who is se

Apr 25, 2019 09:11 PM User ID

What if I've changed my mind? →

Additional Information

Risk Level

● Low
○ Medium
○ High

Risk Type(Select All That Apply)

☑ Nudity/Porn
☐ Sexual Messages
☐ Harassment
☐ Hate Speech
☐ Violence/Threat
☑ Sale or promotion of illegal activities
☐ Self-injury
☐ Other (Please Specify)

Done

Additional Information for your Conversation with [username]

Please describe why this conversation made you or someone else feel uncomfortable or unsafe.

How do you know [username]

I know her/him in real life.
I know her/him online.
I do not know her/him.

What is your relationship with [username]

Family
Your Significant Other
Friend
Acquaintance
Stranger (I don't know them)
Other (Please Specify)

What if I've changed my mind? →

Fig. 2. Screenshot of (a) Participant Conversation Selection Screenshot (b) Participant Messages Risk-flagging Screenshot.

This scale examines the relationship between the use of Facebook, and the formation and maintenance of social capital including bonding, bridging, and maintained social capital [11]. We also utilized Social Media Disorder Scale by van den Eijinden [37] which is a psychometrically sound instrument to measure social media addiction.

2.3.2 Negative Online Experiences. Next, we asked questions regarding potentially negative experiences that participants had on Instagram. Cyberaggression and Cybervictimization (CAV) Scale [33] by Shapka and Maghsoudi was used to measure cyberbullying experience both as a victim and a perpetrator. We also used questions from the Deception of Cyberbullying Victimization and Perpetuation scale by Doane [12] to understand if participants experienced deception and lying on Instagram. Youth Internet Safety Survey (YISS) Unwanted Online Experiences by Mitchel et al. [28] was used to measure sexual solicitation, unwanted exposure to sexual material, and produced sexual images.

2.3.3 Personal Experiences and Demographics. We utilized The Short Warwick-Edinburgh Mental Well-being Scale (SWEMWBS) by Tennant et al. [36] to measure well-being and mental-health, UCLA Loneliness Scale by Hays et al. [19], Patient Health Questionnaire (PHQ-9) [26] to measure depression, Inventory of Statements About Self-injury (ISAS) by Klonsky and Glenn [25] to comprehensively assess the functions of non-suicidal self-injury (NSSI), and Risky Behavior Questionnaire for Adolescents (RBQ-A) by Auerbach and Gardiner [4] to assess risky behavior engagement, impulsiveness, maladaptive coping, risky behavior engagement, and self-esteem of participants. Lastly, we asked demographic questions from participants about their gender, age, location, race, sexual orientation, relationship status, and the caregivers of their teenage years.

2.4 Ground Truth Annotations by Participants

Participants were asked to log in to their primary Instagram account to request a download of their Instagram data file in the form of JSON files in a .zip archive. Once they received their Instagram data file, they were asked to upload the file to our system. Once uploaded, we presented their Instagram private message conversations in a sequential fashion, so they could review their interactions and flag each conversation as 'safe' or 'unsafe', displayed in Figure 2(a). We allowed participants to self-assess the situations that felt risky to them rather than limiting their responses to a predefined subset of risks. Next, participants were asked to provide more details about each risky conversation by first selecting a risk type and then a risk level for each message as shown in Figure 2(b). Drawing on a set of pre-defined risk types derived in a domain-driven manner from existing Instagram reporting feature risk categories [1], we explained to participants that unsafe or uncomfortable interactions may include but were not limited to:

- **Nudity/porn:** Photos or videos of a nude or partially nude people or person.
- **Sexual messages or Solicitations:** Sending or receiving a sexual message ("Sexting") – being asked to send a sexual message, revealing, or naked photo.
- **Harassment:** Messages that contain credible threats, aim to degrade or shame someone, contain personal information to blackmail or harass someone, or threaten to post nude photos of someone.
- **Hate speech:** Messages that encourage violence or attack anyone based on who they are; specific threats of physical harm, theft, or vandalism.
- **Violence/Threat of violence:** Messages, photos or videos of extreme violence, or that encourage violence or attacks anyone based on their religious, ethnic or sexual background.
- **Sale or promotion of illegal activities:** Messages promoting the use, or distributing illegal material such as drugs.
- **Self-injury:** Messages promoting self-injury, which includes suicidal thoughts, cutting, and/or eating disorders.
- **Other:** Other situations that could potentially lead to emotional or physical harm.

We then grounded risk levels based in the existing adolescent online risk literature [39] which operationalized the risk level for youth for how much it is likely to cause emotional or physical harm to them or others: **Low Risk** comprised messages that made the participant uncomfortable but was unlikely to cause emotional or physical harm. **Medium Risk** included messaging which if continued/escalated, would have been likely to cause emotional/physical harm. **High Risk** comprised messages that were deemed dangerous and caused emotional or physical harm to the participant. Participants were additionally asked to provide context for each conversation around why it made them or someone else feel unsafe and the relationships between involved parties; for instance, if the other party in the conversation was an acquaintance, a friend, a boyfriend/girlfriend, or a stranger (ref. right side of the Figure 2(b)). Since pre-existing relationships are known to impact responses in online sexual experience incidents, we considered the knowledge of this relationship relevant to these risk situations [30].

2.5 System Technical Details

Figure 3 illustrates the Instagram Data Donation system architecture. We leveraged several Amazon Web Services (AWS) and other contemporary technologies to develop this system:

- **AWS Relational Database Service (RDS):** was used to save user information and conversations securely in a password-protected MySQL database.
- **AWS Elastic Compute Cloud (EC2):** was created to host and handle the system components which includes the dynamic information flow between the web-front (users input) and the PHP back-end code (handling Database queries or sending Instagram folders to AWS S3 buckets).
- **AWS Simple Storage Service (S3):** was used for data storage for Instagram data folders with restricted access.
- **AWS Lambda:** was used to automatically allocate resources to run codes to power the system back-end and securely process participants' direct messages and media files. The lambda function code (Python) was triggered everytime a new folder was uploaded to the AWS S3 bucket.

[1] https://www.facebook.com/help/instagram/192435014247952

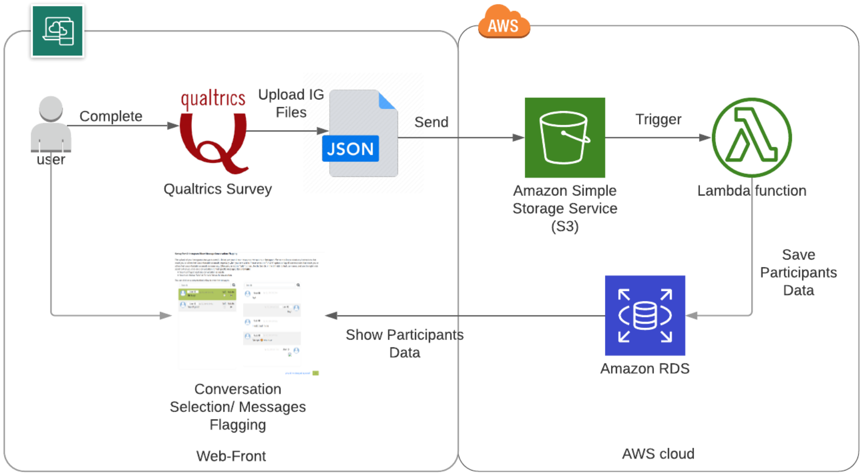


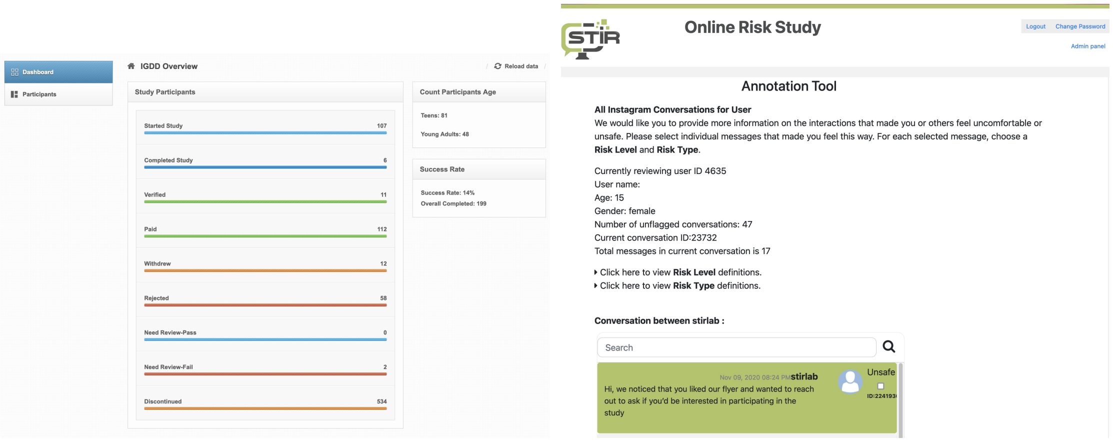
Fig. 3. Instagram Data Donation System Architecture.

- **AWS Simple Email Service (SES):** was used to send participants automatic emails to remind them to complete the study and to confirm successful completion.

We connected the Qualtrics survey to our website by passing variables such as participant ID. After a Qualtrics survey is completed by a participant, the system redirects the participants to a page to upload their Instagram file. The upload page sends the uploaded Instagram folder to be stored in the AWS S3 bucket. Then S3 triggers the lambda function to process the Instagram JSON file, which includes the messages and media files and store the processed data in the RDS MySQL database. After the data file is successfully processed and saved in the database, the conversation selection page retrieves the conversations from the database and displays them to the user to select safe/unsafe conversations and flag the messages of the unsafe conversations based on the risk type and level. Participants are allowed to leave the study at any time and come back to it by leveraging cookies that store participants' progress. If a participant closed the browser at any time during the study, we included a capability to email them the link to continue the survey using AWS SES. After participants successfully completed the study, they received a confirmation email for their completion.

2.5.1 Security Audit. Our technical implementation of the system went through our institutional security audit. We made sure that our system passed all security standards and policies of our institution. Since our data falls into Restricted data according to the university's Data Policy, we made sure to only store data on services (AWS) that are approved by the university. We executed security assessments on the EC2 instance and other services using AWS Inspector. We investigated the Common Vulnerabilities and Exposures and fixed any outstanding issues. Some of the work that we have done for the security of the website is listed below:

- Our AWS is under our an institutional account to be compliant with our institutional contract. Any dependency from external web servers and providers was removed.
- We made sure that all transitions are encrypted including RDS at-rest and in-transit encryption, AWS lambda environment variables encryption, AWS Elastic Block Store (EBS) volume on the EC2 instance encryption, and S3 data 256 AES Server-side encryption. We created a backup plan in AWS Backup that would work for EC2 and RDS.



IGDD Overview

Study Participants

Started Study 107

Completed Study 6

Verified 11

Paid 112

Withdrew 12

Rejected 58

Need Review-Pass 0

Need Review-Fail 2

Discontinued 534

Count Participants Age

Teens: 81

Young Adults: 48

Success Rate

Success Rate: 14%

Overall Completed: 199

Online Risk Study

Annotation Tool

All Instagram Conversations for User
We would like you to provide more information on the interactions that made you or others feel uncomfortable or unsafe. Please select individual messages that made you feel this way. For each selected message, choose a **Risk Level** and **Risk Type**.

Currently reviewing user ID 4635
User name:
Age: 15
Gender: female
Number of unflagged conversations: 47
Current conversation ID:23732
Total messages in current conversation is 17

- Click here to view **Risk Level** definitions.
- Click here to view **Risk Type** definitions.

Conversation between stirlab :

Nov 09, 2020 08:24 PM**stirlab**
Hi, we noticed that you liked our flyer and wanted to reach out to ask if you'd be interested in participating in the study

Unsafe □
ID:224193

Fig. 4. Screenshot of (a) Participants Dashboard (b) Annotation Tool.

- Any connection to / from EC2 server and between the server and other services like RDS (Database) and lambda function uses the Secure Sockets Layer (SSL). We made sure all the instances are updated to most available versions.

2.6 Data Ground Truth and Annotation Tool

To make sure that all the data is annotated for risks with consistently high quality, two research assistants are employed to review each conversation to identify potential risks that were missed, as third party annotators. Once the annotation by the two research assistants is complete, we calculate inter-rater reliability for two coders for each conversation. If two out of three coders (including participants) agree on specific risk instances, then we can reliably call that instance a risk. We developed a web-based tool (Figure 4(b)) to facilitate this annotation process. The third party annotators were provided a similar interface to participants' interface to flag any unsafe messages.

2.7 Data Verification Process

To keep track of number of participants who participated in the study and ease the process of the data verification we developed a web-tool displayed on Figure 1. At the time of writing of this paper, 107 youth had started the study and were in route to completing the study. In addition, 123 participants completed the study and passed the data verification quality check. We adopted various quality check to make sure participants were not answering the survey questions arbitrarily, were genuine in their responses, and completed the study attentively. We made sure that participants met the eligibility criteria such as having at least 15 conversations and had a history of Instagram for the duration specified in our inclusion criteria and at least 2 unsafe conversations with exchanged messages. We removed participants who did not answer attention check survey questions (e.g., Select "Strongly Agree" for this item) or two independent age verification questions correctly, or who took unrealistically little time for completion. Checked the quality of their Instagram data file to make sure it was from a real youth participant and not from a fake or bot account. Our recruitment efforts happened during the COVID-19 pandemic, which presented new complexities, since we could not recruit in person. e.g., it slowed data collection and resulted in some participants failing quality checks.

2.8 Participants Demographics and Dataset Characteristics

To collect data belonging to individuals from varied demography within the US we promoted our study on social media especially Facebook and Instagram. We did not limit our recruitment process online and also contacted more than 650 youth-serving organizations such as suicide prevention programs, group homes, and early pregnancy centers offline. We also affirmed that our data is aligned with published data from the United States Census site [1]. Here are some descriptive statistics of our collected data: From 123 verified participants 70% are females, 21% males and 9% non-binary or prefer to self-identify individuals, all ranging between the ages of 13 and 21 years. (Average Age = 15, Standard Dev. = 5.5. Most of the participants recognized themselves as heterosexual or straight 48%, however, our dataset also includes 29% bisexual, 11% homosexual, and 13% who chose to self identify. Next, we found that 39% of our participants are White, 20% Black/African-American, 16% Asian or Pacific Islander, and 7% Hispanic/Latino and 16% belonging to mixed races or who preferred not to self-identify. From the 123 verified participants we collected 22,477 conversations, where 1,789 conversations were labeled as unsafe by the respective participants. The total number of messages included in these conversations is over 5 million, out of which 5,824 of them were flagged by participants for risk type and levels. On average 60% of the messages were flagged as low, 26% as medium and 14% as high risk levels.

3 FINDINGS: LESSONS LEARNED

While conducting our research, we overcame several challenges ranging from technical issues, dealing with gathering a sensitive dataset, to ethical considerations that we share with the research community.

3.1 Overcoming Technical Challenges

We developed our data collection system as following the Cambridge Analytica data breach [21], after which Facebook services for providing data to researchers were discontinued. Consequently, after the launching of the IGDD system, multiple technical challenges appeared that required our developers to resolve these issues in an efficient manner.

3.1.1 Leveraging AWS Services. One of the challenges was that Instagram changed the users' folder organization and JSON format multiple times after launching the study. Once we realized there was a major change in the file format, we shut down the production server and directed participants to a maintenance page to let them know that the study is still available and we reached out to them to proceed once the issue is fixed. Instead of using the front-end pages to test the new code, AWS offered an integrated development environment called Cloud 9 to test the new code faster. In addition, AWS provided flexible integration to new services to the lambda function. For example, processing the images and videos caused a performance bottleneck; therefore, we used Simple Queue Service (SQS) to enhance processing. By integrating the SQS to the lambda function, we were able to process multiple media files at the same time.

3.1.2 User-Centered System. In addition to the security and efficiency of the system, user experience was another critical focus of our system as youth were the target users. While most were able to complete the survey part with no particular difficulty; however, many expressed confusion when uploading their data. We made sure that the error messages were precise and clear. We also looked at how we could adapt our system to handle users' common mistakes automatically. Specifically, we established an FAQ page which described the common issues that a participant could face during the study. Most of the issues were related to uploading their Instagram data, as it could be very large in size or in different formats. We also created a systematic approach to resolve any issue that stopped the upload process. We had research assistants to resolve these issues and follow up with participants in a timely manner. .

3.2 Overcoming Privacy and Ethical Challenges

Collecting social media data is a sensitive subject itself [15], and when the data is collected from minors the difficulties and precautions required increase drastically. Therefore, preserving the confidentiality and privacy of the participants becomes very important, considering the complexity and the sensitive nature of our private dataset compared to public datasets. Apart from obtaining IRB approval for the study, we adopted a series of measures to ensure that the participants were protected and the data gathering process proceeded in an ethical way.

3.2.1 Legal and Ethical Challenges. We disclosed ourselves as mandated child abuse reporters [6] for urgent cases of risk posed to minors. As mandated reporters, if we were to have reasonable suspicion that a child has been abused, neglected or threatened of harm in the state, we were required to contact the Florida Abuse Hotline to report the incident. The Hotline counselor would determine if the information provided met legal requirements to accept a report for investigation. We clearly stated our federal obligations to report any child pornography to authorities. Consequently, we explicitly warned against uploading any digital content containing nudity of minors. To assist the participants, we gave detailed instructions on how to remove such data before uploading it to our server. In any exceptional cases that any clear child pornography was found by researchers, several steps will be taken for a proper report.

3.2.2 Privacy, Data Protection, and Sharing. To protect the privacy of our participants and prevent subpoena of data, we obtained a National Institute of Health (NIH) Certificate of Confidentiality. For publications resulting from this dataset in the future, we considered different de-identification measures. We settled on removing any personally identifiable information from textual or image data, including paraphrasing or editing the content of any presented data, based on guidance in prior research. Due to the sensitive nature of the dataset, it will not be made publicly available for use, but maybe shared as a restricted dataset. Individuals requesting third-party access to the more sensitive raw data of teen social media data (de-identified within reasonable standards using automated de-identification tools) will need to show an established record of relevant, published research to validate why they should have access to this data, IRB approval from their home institutions, in addition to meeting the requirements for reuse and redistribution as described in the IRB protocol. For the distribution of more sensitive teen social media data, individuals requesting third party access will sign a data use agreement reviewed by applicable institutional departments. For having a clear timline on how long we would keep the data, we made sure to follow the university's, state, and NSF data retention policies.

3.2.3 Additional Safety Precautions. All researchers completed the IRB Human Subjects CITI training and UCF's Youth Protection Program training to ensure the safety of our participants. They were prohibited from downloading the data on personal devices. All students who helped verify and annotate the donated data from the participants were given adequate breaks and mental health support, given that some of the risky behaviors presented in the data could be traumatizing. As researchers we were unable to make diagnostic clinical decisions about a participant's mental health, but we provided participants help and support resources in case they needed it. These resources included Mental Health Resources (*https://www.adolescenthealth.org/Resources/Clinical-Care-Resources/Mental-Health/*), Crisis Intervention Resources (*https://www.crisistextline.org/*), Trevor Lifeline (*https://www.thetrevorproject.org/get-help-now/*), Suicide Prevention Resources (*https://suicidepreventionlifeline.org*), and Child Abuse Hotline (*https://www.childhelp.org/hotline/*).

4 DISCUSSION: LIMITATIONS AND FUTURE RESEARCH

Our work embodies a foundation to online risk detection by creating a human-centered ecologically valid dataset that, as far to our knowledge, is unprecedented. The self-reported annotations of youth who have been exposed to online risk shine a light on the perspectives of the victims. The private conversations between the perpetrators and the victims would additionally be a valuable source to establish the ground truth for detecting unsafe incidents. Next, the wide range of online risk annotations spanning across textual and image data introduces the opportunity for the development of multimodal risk detection systems that could be provided as timely and scalable solutions to provide support for current and potential victims of online harm. It should be noted that dealing with such sensitive data is accompanied by the various challenges that researchers should carefully address. Privacy protection, and ethical usage of private data including the transparency and interpretability of the results should be the utmost priority. Such consideration should also extend to the speculated usage of the applications when deployed in real-life scenarios. Metrics for evaluating the performance of the models build on such data should be aligned with human-centered perspectives to incorporate the potential impact on the users as well as any negative consequences.

A key strength of our work is that we collected a dataset of private Instagram conversations from youth. One limitation of our work relates to difficulties with reproducibility of the results from this private dataset. Because of the sensitivity of the dataset, we are unable to share it publicly. However, we are willing to collaborate and share part of the dataset with researchers from accredited institutions. Also, we cannot use any cloud based APIs for the analysis of this data so as to not reveal any data to third parties. In addition, our research is based on Instagram, which has its own platform affordances. Therefore, to generalize the results produced from this data, researchers will need to investigate private data from other platforms. Our data collection tool was created by keeping the Instagram platform in mind, and the data processing pipeline was based on how data is organized by Instagram. We believe the general architecture of our tool could be tailored to other social media platforms. Finally, a unique strength but also a limitation of our study is that participants' labels for unsafe conversations is dependent on their perspective of risks, thus incorporating subjectivity. Participants' labels provide us more understanding on how and why a conversation was labeled by the participant as risky. To overcome this limitation that the labeling is solely from the perspective of the participant, we are also in the process of having research assistants to review and annotate our dataset. These researchers will identify potential risks that were missed and gain qualitative insights into the private digital lives of youth. Future work could include conducting post-hoc follow-up interviews with the youth who participated to understand how participants felt about reviewing and flagging their past risk experiences. The goal could be to have them reflect on their past experience to evaluate how they felt about the interface and provide implications for design and best mental health practices for protecting them accordingly. Complementarily, follow-up interviews with the third party annotators of the data in the future can help to understand their thoughts during the annotation process and its effect on their well-being. Taken together, this case study paves the way for further research on crafting ethical methodologies of sensitive social media data collection that are sensitive to the needs and demands of different stakeholders.

ACKNOWLEDGMENTS

This research is supported by the U.S. National Science Foundation under grants #IIP-1827700 and by the William T. Grant Foundation grant #187941. 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. Special thanks to all the participants who donated their data and research assistants who assisted with the data collection and annotations. We would also like to thank Zinan Zhang, Jessica Medina, and Marcos Chacon for creating the video for this case study.

REFERENCES

[1] [n.d.]. U.S. Census Bureau QuickFacts: United States. https://www.census.gov/quickfacts/fact/table/US/PST045219

[2] Alaa Althubaiti. 2016. Information bias in health research: definition, pitfalls, and adjustment methods. *Journal of multidisciplinary healthcare* 9 (2016), 211.

[3] Monica Anderson and Jingjing Jiang. 2018. Teens, Social Media & Technology 2018 | Pew Research Center. http://www.pewinternet.org/2018/05/31/teens-social-media-technology-2018/

[4] Randy P Auerbach and Casey K Gardiner. 2012. Moving beyond the trait conceptualization of self-esteem: The prospective effect of impulsiveness, coping, and risky behavior engagement. *Behaviour research and therapy* 50, 10 (2012), 596–603.

[5] Karla Badillo-Urquiola, Afsaneh Razi, Jan Edwards, and Pamela Wisniewski. 2020. Children's Perspectives on Human Sex Trafficking Prevention Education. In *Companion of the 2020 ACM International Conference on Supporting Group Work*. 123–126.

[6] 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* 4, CSCW3 (2021), 1–46.

[7] Anna Costanza Baldry, David P Farrington, and Anna Sorrentino. 2017. School bullying and cyberbullying among boys and girls: Roles and overlap. *Journal of Aggression, Maltreatment & Trauma* 26, 9 (2017), 937–951.

[8] Fanni Bányai, Ágnes Zsila, Orsolya Király, Aniko Maraz, Zsuzsanna Elekes, Mark D Griffiths, Cecilie Schou Andreassen, and Zsolt Demetrovics. 2017. Problematic social media use: Results from a large-scale nationally representative adolescent sample. *PloS one* 12, 1 (2017), e0169839.

[9] Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics* 6 (2018), 587–604.

[10] Ruha Benjamin. 2019. Assessing risk, automating racism. *Science* 366, 6464 (2019), 421–422.

[11] Pierre Bourdieu and Loïc JD Wacquant. 1992. *An invitation to reflexive sociology*. University of Chicago press.

[12] Ashley N Doane, Michelle L Kelley, Evelyn S Chiang, and Miguel A Padilla. 2013. Development of the cyberbullying experiences survey. *Emerging Adulthood* 1, 3 (2013), 207–218.

[13] Nicole B. Ellison, Charles Steinfield, and Cliff Lampe. 2007. The Benefits of Facebook "Friends:" Social Capital and College Students' Use of Online Social Network Sites. *Journal of Computer-Mediated Communication* 12, 4 (July 2007), 1143–1168. https://doi.org/10.1111/j.1083-6101.2007.00367.x

[14] Sindhu Kiranmai Ernala, Michael L Birnbaum, Kristin A Candan, Asra F Rizvi, William A Sterling, John M Kane, and Munmun De Choudhury. 2019. Methodological gaps in predicting mental health states from social media: triangulating diagnostic signals. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–16.

[15] Casey Fiesler and Nicholas Proferes. 2018. "Participant" perceptions of Twitter research ethics. *Social Media+ Society* 4, 1 (2018), 2056305118763366.

[16] General Data Protection Regulation (GDPR). 2021. Art. 20 GDPR – Right to data portability | General Data Protection Regulation (GDPR). https://gdpr-info.eu/art-20-gdpr/

[17] Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. 2020. Towards a critical race methodology in algorithmic fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 501–512.

[18] Naeemul Hassan, Amrit Poudel, Jason Hale, Claire Hubacek, Khandaker Tasnim Huq, Shubhra Kanti Karmaker Santu, and Syed Ishtiaque Ahmed. 2020. Towards Automated Sexual Violence Report Tracking. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14. 250–259.

[19] Ron D Hays and M Robin DiMatteo. 1987. A short-form measure of loneliness. *Journal of personality assessment* 51, 1 (1987), 69–81.

[20] Nicola Henry and Anastasia Powell. 2018. Technology-facilitated sexual violence: A literature review of empirical research. *Trauma, violence, & abuse* 19, 2 (2018), 195–208.

[21] Jim Isaak and Mina J Hanna. 2018. User data privacy: Facebook, Cambridge Analytica, and privacy protection. *Computer* 51, 8 (2018), 56–59.

[22] Sweta Karlekar and Mohit Bansal. 2018. Safecity: Understanding diverse forms of sexual harassment personal stories. *arXiv preprint arXiv:1809.04739* (2018).

[23] Seunghyun Kim, Afsaneh Razi, Gianluca Stringhini, Pamela Wisniewski, and Munmun De Choudhury. 2021. A Human-Centered Systematic Literature Review of Cyberbullying Detection Algorithms. (2021).

[24] Seunghyun Kim, Afsaneh Razi, Gianluca Stringhini, Pamela Wisniewski, and Munmun De Choudhury. 2021. You Don't Know How I Feel: Insider-Outsider Perspective Gaps in Cyberbullying Risk Detection. In *Proceedings of the International AAAI Conference on Web and Social Media.*

[25] E David Klonsky and Catherine R Glenn. 2009. Assessing the functions of non-suicidal self-injury: Psychometric properties of the Inventory of Statements About Self-injury (ISAS). *Journal of psychopathology and behavioral assessment* 31, 3 (2009), 215–219.

[26] Kurt Kroenke and Robert L Spitzer. 2002. The PHQ-9: a new depression diagnostic and severity measure.

[27] Suzet Tanya Lereya, Catherine Winsper, Jon Heron, Glyn Lewis, David Gunnell, Helen L Fisher, and Dieter Wolke. 2013. Being bullied during childhood and the prospective pathways to self-harm in late adolescence. *Journal of the American Academy of Child & Adolescent Psychiatry* 52, 6 (2013), 608–618.

[28] Kimberly J Mitchell and Lisa M Jones. 2011. Youth Internet Safety Study (YISS): Methodology Report. (2011).

[29] Anthony T. Pinter, Pamela J. Wisniewski, Heng Xu, Mary Beth Rosson, and Jack M. Caroll. 2017. Adolescent Online Safety: Moving Beyond Formative Evaluations to Designing Solutions for the Future. In *Proceedings of the 2017 Conference on Interaction Design and Children - IDC '17*. ACM Press, Stanford, California, USA, 352–357. https://doi.org/10.1145/3078072.3079722

[30] Afsaneh Razi, Karla Badillo-Urquiola, and Pamela J. Wisniewski. 2020. Let's Talk about Sext: How Adolescents Seek Support and Advice about Their Online Sexual Experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing

Machinery, Honolulu, HI, USA, 1–13. https://doi.org/10.1145/3313831.3376400

[31] Afsaneh Razi, Seunghyun Kim, Ashwaq Soubai, Gianluca Stringhini, Thamar Solorio, Munmun De Choudhury, and Pamela Wisniewski. 2021. A Human-Centered Systematic Literature Review of the Computational Approaches for Online Sexual Risk Detection. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 465 (Oct. 2021), 38 pages. https://doi.org/10.1145/3479609

[32] Michael J Rosenfeld, Reuben J Thomas, and Sonia Hausen. 2019. Disintermediating your friends: How online dating in the United States displaces other ways of meeting. *Proceedings of the National Academy of Sciences* 116, 36 (2019), 17753–17758.

[33] Jennifer D. Shapka and Rose Maghsoudi. 2017. Examining the validity and reliability of the cyber-aggression and cyber-victimization scale. *Computers in Human Behavior* 69 (April 2017), 10–17. https://doi.org/10.1016/j.chb.2016.12.015

[34] Robert Slonje, Peter K Smith, and Ann Frisén. 2017. Perceived reasons for the negative impact of cyberbullying and traditional bullying. *European journal of developmental psychology* 14, 3 (2017), 295–310.

[35] Ashima Suvarna, Grusha Bhalla, Shailender Kumar, and Ashi Bhardwaj. 2020. Identifying Victim Blaming Language in Discussions about Sexual Assaults on Twitter. In *International Conference on Social Media and Society (SMSociety'20)*. Association for Computing Machinery, Toronto, ON, Canada, 156–163. https://doi.org/10.1145/3400806.3400825

[36] Ruth Tennant, Louise Hiller, Ruth Fishwick, Stephen Platt, Stephen Joseph, Scott Weich, Jane Parkinson, Jenny Secker, and Sarah Stewart-Brown. 2007. The Warwick-Edinburgh mental well-being scale (WEMWBS): development and UK validation. *Health and Quality of life Outcomes* 5, 1 (2007), 1–13.

[37] Regina J. J. M. van den Eijnden, Jeroen S. Lemmens, and Patti M. Valkenburg. 2016. The Social Media Disorder Scale. *Computers in Human Behavior* 61 (Aug. 2016), 478–487. https://doi.org/10.1016/j.chb.2016.03.038

[38] Pamela Wisniewski, Haiyan Jia, Na Wang, Saijing Zheng, Heng Xu, Mary Beth Rosson, and John M Carroll. 2015. Resilience mitigates the negative effects of adolescent internet addiction and online risk exposure. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 4029–4038.

[39] Pamela Wisniewski, Heng Xu, Mary Beth Rosson, Daniel F. Perkins, and John M. Carroll. 2016. Dear Diary: Teens Reflect on Their Weekly Online Risk Experiences. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 3919–3930. https://doi.org/10.1145/2858036.2858317 event-place: San Jose, California, USA.

[40] Jacob O Wobbrock and Julie A Kientz. 2016. Research contributions in human-computer interaction. *interactions* 23, 3 (2016), 38–44.