# Understanding the Digital Lives of Youth: Analyzing Media Shared within Safe Versus Unsafe Private Conversations on Instagram

Shiza Ali
Boston University
Boston, Massachusetts, USA
shiza@bu.edu

Afsaneh Razi
University of Central Florida
Orlando, Florida, USA
afsaneh.razi@knights.ucf.edu

Kim Seunghyun
Georgia Institute of Technology
Atlanta, Georgia, USA
seunghyun.kim@gatech.edu

Ashwaq Alsoubai
University of Central Florida
Orlando, Florida, USA

Joshua Gracie
University of Central Florida
Orlando, Florida, USA
joshua_gracie@knights.ucf.edu

Munmun De Choudhury
Georgia Institute of Technology
Atlanta, Georgia, USA
munmund@gatech.edu

Pamela J. Wisniewski
University of Central Florida
Orlando, Florida, USA
pamela.Wisniewski@ucf.edu

Gianluca Stringhini
Boston University
Boston, Massachusetts, USA
gian@bu.edu

## ABSTRACT

We collected Instagram Direct Messages (DMs) from 100 adolescents and young adults (ages 13-21) who then flagged their own conversations as safe or unsafe. We performed a mixed-method analysis of the media files shared privately in these conversations to gain human-centered insights into the risky interactions experienced by youth. Unsafe conversations ranged from unwanted sexual solicitations to mental health related concerns, and images shared in unsafe conversations tended to be of people and convey negative emotions, while those shared in regular conversations more often conveyed positive emotions and contained objects. Further, unsafe conversations were significantly shorter, suggesting that youth disengaged when they felt unsafe. Our work uncovers salient characteristics of safe and unsafe media shared in private conversations and provides the foundation to develop automated systems for online risk detection and mitigation.

## CCS CONCEPTS

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

## KEYWORDS

Adolescents, Teens, Datasets, Instagram, unsafe private conversations, Image Analysis


*CHI '22, April 29-May 5, 2022, New Orleans, LA, USA*

ACM ISBN 978-1-4503-9157-3/22/04…$15.00
https://doi.org/10.1145/3491102.3501969

**ACM Reference Format:**
Shiza Ali, Afsaneh Razi, Kim Seunghyun, Ashwaq Alsoubai, Joshua Gracie, Munmun De Choudhury, Pamela J. Wisniewski, and Gianluca Stringhini. 2022. Understanding the Digital Lives of Youth: Analyzing Media Shared within Safe Versus Unsafe Private Conversations on Instagram. In *CHI Conference on Human Factors in Computing Systems (CHI '22), April 29-May 5, 2022, New Orleans, LA, USA.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3491102.3501969

## 1 INTRODUCTION

Adolescents and young adults are among the most avid users of social media platforms. For instance, Pew Research [1] reports that 72% of teens use Instagram, making it one of the most popular social media platforms among youth. Given the popularity of Instagram, several researchers [21, 41, 53] have studied the sharing practices of youth on this platform. A line of studies have focused on what types of photos were shared publicly on Instagram compared to other social media platforms such as Snapchat [29], while other studies showed how frequently youth shared Instagram photo posts compared to adults [22]. These studies are valuable as they have contributed to the understanding of how youth engage differently than adults on social media, as well as how these behaviors may differ based on the unique affordances of specific social media platforms. A limitation of these past studies, however, is that they primarily focused on publicly observed interactions rather than private interactions through more intimate channels, such as direct messages.

The public versus private discourse of youth on social media has also become a research topic of great interest, combined with the role privacy plays in the online safety of youth [23, 58, 59]. For instance, Marwick and boyd's work [6, 7, 36] highlighted how teens go to great lengths to "be in public without always being public"(p.1052), while the adolescent online safety literature [33] reveals that youth treat risk-taking as a learning process that shapes their subsequent privacy behaviors. Often, youth take retroactive

protective behaviors (e.g., deleting a picture or blocking someone) after a risky interaction heightens their concern for privacy [23]. Yet, due to the inability of being able to obtain private social media data from youth, few researchers have been able to study these less visible interactions in a meaningful way. Most of the research on online risk experiences (e.g., unwanted sexual solicitations and cyberbullying) of youth has relied heavily on self-reports, rather than in-depth analyses of social media trace datalogs [45].

However, more recent research [15, 26, 27, 45] has shown the value of using social media trace data as a means to unobtrusively study youth online risk experiences at-scale. For instance, Hassan et al. [16] collected and utilized tweets with the #MeToo hashtag that was used by women to disclose their harassment or violence experiences to build an automated sexual violence report tracker. For instance, Kim et al. studied an online peer support platform to build a classifier that detected bullying narratives and showed a significant difference between insider (i.e., self-reports) versus outsider (i.e., third-party annotators) ground truth [27]. They found that third-party annotators are more conservative than victims when classifying a post as bullying or not. These studies demonstrate the importance of incorporating first-person perspectives of their unsafe online interactions and that much can be learned about the online risk experiences of youth through analyzing their social media trace data. Our study is one of the first to go beyond analyzing semi-public discourse to collect and examine *private* Instagram conversations from youth specific to their unsafe online interactions with others. To do this, we recruited 100 adolescents and young adults (ages 13-21) to share their Instagram data with us (obtaining parental consent for those who were minors) through a secure web-based system. We collected 4,752,560 Direct Messages (DMs) within 11,062 private conversations. Then, we had each participant flag their own direct message conversations that made them or someone else feel uncomfortable or unsafe. A total of 1,452 (13.13%) conversations were flagged by participants and included experiences of harassment, sexual solicitations, violence, self-injury, etc. Leveraging this rich, risk-flagged data, we conducted quantitative analyses to identify key differences between the multimodal data (i.e., number of messages, textual content, and images) shared in safe versus unsafe conversations. Through this analysis, we address the following high-level research questions:

- **RQ1 – *Between-group Differences:*** How does the use of media vary between safe and unsafe private conversations?
- **RQ2 – *Image Characteristics:*** What are the characteristics of images shared in safe/unsafe conversations?

Compared to other mainstream social media platforms like Facebook and Twitter, Instagram is predominantly a video and photo sharing platform where users can share media both privately and publicly. Therefore, our paper primarily analyzes media (e.g., images, links) shared in the private domain, i.e., through direct messaging. For RQ1, we found that unsafe conversations contain significantly fewer messages and media (e.g., images and links), indicating that unsafe-flagged conversations are mostly one-sided with our participants being the receivers of unsafe DMs. For RQ2, we found that images shared within unsafe conversations contained significantly more people, while safe conversations contained more object-based photos. We also uncovered that screenshot images were shared in conversations with acquaintances, friends, or family. Interestingly, quite a few of the screenshots shared within safe conversations were of unsafe conversations participants had with others, indicating that private DMs may be an important means for disclosing and getting support from trusted friends related to these negative interactions with others online. This research makes the following empirical contributions to the fields of Human-Computer Interaction (HCI) specific to the literature on adolescent online safety and risks:

- We collected a rich and difficult-to-obtain dataset of private social media conversations youth had with others. Importantly, we had participants flag their own conversations as safe or unsafe, which provides insight into the first-person perspectives regarding online safety of youth.
- We identified key differences between safe and unsafe conversations, such as the number of messages exchanged and media shared. These findings shed light on how youth engage and disengage when risk is perceived.
- We developed a method to separate regular images from screenshots to better understand the kind of images shared in private conversations and characterized images within unsafe conversations to inform future research on automated risk detection specific to the private online interactions of youth.

## 2 RELATED WORK

We review previous research on adolescent online safety, utilizing social media data to understand users and their behaviors, and how automated approaches were developed to detect online risks.

### 2.1 Youth, Social Media, and Online Safety

The Internet and social media provide great opportunities for youth to learn, but also exposes them to various online risks, including sexual predation, cyberbullying, and mental-health issues [33]. Adolescent online safety has become an established research area, and researchers have employed various empirical methods to study youth online risk behaviors [45]. Yet, most of our knowledge about what youth are doing online, as well as the outcomes associated with these online activities, is derived from large-scale surveys [24], diary studies [38] or interview studies [57] that ask teens to self-report on their online experiences. As such, Pinter et al.'s review of the online safety literature identified the need for more empirical methods that go beyond self-reports to document teens' unfiltered online risk experiences [45], as self-report methods such as surveys or interviews are prone to recall bias [11].

Furthermore, much of the literature on youth and their online safety uses a privacy-focused lens to scrutinize the personal information teens disclose publicly, which increases the likelihood of unsafe or negative online interactions [4, 36]. Recently, researchers have begun studying the online trace data from youth to move beyond self-reported data to understand adolescents' online risk behaviors [14, 15, 27, 46]. These studies analyzed public or semi-public online and social media interactions of youth around their online risk behavior. For instance, Razi et al. [46] conducted a thematic analysis of public posts by adolescents on a peer-based mental

health support platform to understand their support seeking behaviors for online sexual experiences. They found that adolescents often received unwanted nudes from strangers and struggled with how to turn down sexting requests from people they knew and while seeking support on the platform they received unwanted sexual solicitations. Meanwhile, researchers have yet to unlock the mystery of how teens engage in non-public online forums. One study examined semi-public social media behaviors of youth on Facebook by friending them and found that a quarter of the profiles contained sexual and romantic references [8]. Our work builds upon this literature by examining youths' private social media interactions to comprehensively understand online risk behavior. As such, we contribute to the literature by collecting and examining a large corpus of private social media data from youth to uncover more insights into their online safety and risks from their first-person perspective.

## 2.2 Analyzing Social Media Trace Data

Social Computing research communities have a long-standing history of analyzing social media trace data to understand users and their behaviors. For instance, Hu et al. [19] used the Instagram API to identify different profiles and popular photo categories based on their dataset. They found that users on Instagram have distinct characteristics in terms of the photo they share. For example, there exists "selfies-lovers," "food-lovers," "pet-lovers" etc. Furthermore, several researchers have analyzed human sentiments from images based on both image features and contextual social network information [30, 56, 62]. These works found that textual information can provide semantic meanings and sentiment predictions for images to help us better understand user behavior. Previous research has also been conducted to analyze the photo sharing practices of youth on Instagram. For example, Jang et al. [22] concluded that teens interact (like or comment) more with images, yet share fewer photos than adults. Youth are also more likely to remove photos based on the number of Likes received. Additionally, teens tend to post about same topics, whereas adults post more diverse content, such as world travel, as well as photos containing more and different people. An online survey focusing on "photo-elicitation" practices between Snapchat and Instagram showed that youth share more polished photos on Instagram, while they are more willing to share less "picture-perfect" content on Snapchat, which allows for more ephemeral sharing [29].

Our research builds upon these previous studies by analyzing the social media trace data of youth specific to their online risk experiences on Instagram. Our work extends the prior work in several ways: 1) We accomplished the arduous task of designing an IRB-approved study to collect Instagram private messages directly from adolescents and young adults (ages 13-21), 2) We had participants flag their own conversations as safe/unsafe, and 3) We examine images in shared in DMs on social media data to ascertain key differences based on risk. In the next section, we show how this is foundational and prerequisite work towards the goal of building robust machine learning classifiers for online risk detection that are tailored to youth.

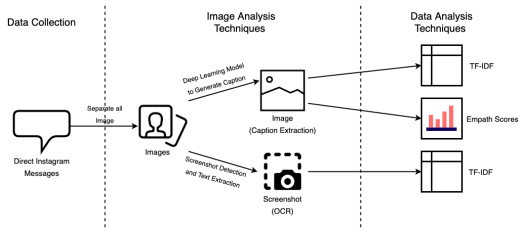


**Figure 1: Our data collection and analysis pipeline.**

## 2.3 Grounding Automated Risk Detection Approaches in the Lived Experience of Youth

Researchers have examined ways to build robust systems focused on detecting specific online risks. Such detection algorithms are mostly designed and developed with machine learning and computational social scientists working closely [27]. The studies in this research area aim to identify characteristics for the pertinent online risk and train a machine learning model that classifies risk incidents based on data from social media. For example, one study [17] detects incidents of cyberbullying over images in Instagram while another [32] looked at Instagram comments under public images to predict whether future comments will contain hostility based on linguistic and social features from earlier comments. While the domain of automated risk detection is quite mature, the past literature often falls short of establishing a robust ground truth and utilizing an ecologically valid dataset in the development of these automated risk detection models [40, 49, 54]. To address the challenge of obtaining a publicly available dataset with ground truth annotations [49, 50], researchers have utilized crowd-sourcing platforms [42]. The ground truth obtained through such approach leaves critical questions to the researchers such as how researchers control for the different perceptions of cyberbullying based on the subjective nature of the experience of being cyber-bullied [9] as well as the diversity of stakeholders [39]. To date, computational researchers building risk detection classifiers have relied heavily on the publicly available posts [28, 47]. In the case of sexual risk detection, scrapping public posts from social media platforms such as Twitter have been used for identifying sexual assault victim-blaming language [52]. Considering that people have different levels of comfort depending on the target audience of their posts [34], the datasets based on posts that are open to the public, and moreover the automated risk detection models built on them, could have little, if any, application to real-world scenarios. By examining conversation data in DMs of Instagram users, we aim to analyze the intimate patterns between safe and unsafe conversations with self-reported annotations of our dataset.

# 3 DATA AND METHODS

In this section, we describe how we collected our data, the risk-flagging process, and the quantitative methods approach employed to analyze this large corpus of social media data. Figure 1 shows the process we used to collect our data, process the images and perform data analysis.

### 3.1 Data Collection Approach

Following approval from the Institutional Review Boards (IRBs) of the authors' institutions, we obtained informed consent from eligible participants over the age of 18; for those under 18, we obtained informed consent from their parents followed by their informed assent. We recruited participants between the ages of 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. We explained to participants that unsafe or uncomfortable interactions may include but not limited to the following categories/types, identified in a domain-driven manner, grounded in the existing adolescent online risk literature [61] and the existing risk categories from Instagram reporting feature aligned with what participants experience on Instagram:

- **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 encouraging or 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 selected Instagram as the platform for data collection given its popularity among youth and young adults [2]. For instance, Instagram is one of the top social media platforms being used by teens in the U.S. between the ages 13 to 17 [2]. Due to General Data Protection Regulation (GDPR), social media users have the right to download their personal data [12] for their own use and to share it without restriction; therefore, Instagram provides a means for users to download their personal data, which includes private, direct message conversations co-owned by others. As such, participants were asked to login to their primary Instagram account to request a download of their Instagram data file in the form of JSON files in a .zip archive. Therefore, we developed a secure web-based system leveraging Amazon Web Services, RDS, EC2, PHP, Python, and other technologies to create a social media data collection system. Participants were asked to request their data file from Instagram, upload it to our system, and subsequently view their private message conversations to flag them.

### 3.2 Risk-Flagging and Data Verification Process

Once participants successfully uploaded their Instagram data file, we presented their Instagram private message conversations in reverse chronological order, so they could review their past interactions and flag each conversation as 'safe' or 'unsafe.' We allowed participants to self-assess the situations that felt unsafe or uncomfortable to them rather than limiting their responses to a predefined subset of risks. Once they flagged each conversation, participants were asked to provide more context details about each unsafe interaction, for instance, we asked them to "describe why this conversation made you or someone else feel uncomfortable or unsafe."

Upon completing the study, a team of researchers verified the data and compensated participants who passed quality checks with a $50 Amazon gift card for their data and time. We included several quality checks questions to make sure participants answered the questions attentively and provided a real data file. The data verification team checked the following items to make sure that the participants were genuine and provided good quality data, filtering out any data that had low quality or seemed fictitious:

- Checked the time that it took participants to complete the study to remove participants who took unrealistically little time for completing the study.
- Made sure that participants met the eligibility criteria such as having at least 15 conversations on Instagram and had a history of Instagram for the duration specified in our inclusion criteria.
- Checked the details of the unsafe conversations to make sure participants had at least two unsafe conversations, to filter out the single message by bots or strangers that are not relevant to this study.
- 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.
- 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. This included reviewing all private conversations included in the file to ensure there was a sufficient history of past conversations, indicating that participants did not fabricate any of the data.

### 3.3 Ethical Considerations

Due to the complex and sensitive nature of the dataset, we took the utmost care to preserve the confidentiality and privacy of the participants. Following the recommendations of Badillo-Urquiola et al. [3] on conducting risky research with minors, in addition to obtaining IRB approval for our study, we disclosed our status as mandated child abuse reporters in the case of imminent risk posed to a minor. We also explained our federal obligation to report child pornography to the proper authorities and gave explicit warnings not to upload digital imagery depicting nudity of a minor. We gave instructions for how to remove such media from the data. Additionally, we obtained a National Institute of Health Certificate of Confidentiality, which further ensures participant privacy and prevents the subpoena of the data during the legal discovery.

We also took special care regarding to data and analyses presented in the paper. For instance, we removed all personally identifiable information in any textual or image data reported in our results, paraphrased all quotations, and recreated privately shared images to ensure the confidentiality of our participants and all other individuals who participated in direct message conversations with our participants. For images that were publicly available, such as memes, we did not change them. We also chose not to use any cloud-based services (e.g., Google Vision API) when analyzing our data to avoid sharing the data with third-parties. Researchers analyzing the data were required to complete IRB Human Subjects CITI training and not permitted to download the data on any personal devices. We also provided mental health support and adequate breaks for students who helped verify and qualitatively analyze the data as some of the content could be triggering or explicit.

## 3.4 Participant Recruitment and Demographics

Our goal was to recruit a wide variety of youth from diverse backgrounds. We accomplished this through contacting more than 650 youth-serving organizations, particularly those who work with at-risk youth, suicide prevention programs, group homes, LGBTQAI+ centers, and early pregnancy centers. We also posted and promoted our study on Facebook and Instagram to get a wider audience across US. We checked the demographic distribution of our participants with published data from the United States Census site by government [5], and they are alighted with those data. The verified participants ($N$=100) in our study were between 13 and 21 years old, with the average age being 16 years old (std=2.03). Figure 2 displays the frequency of the number of participants and their ages. Most of the participants in our were female (68%), with 24% from males, and 8% from non-binary or individuals that preferred not to answer. Participants' race distribution is as follows: 41% White, 19% Black/African-American, 16% mixed races or preferred to self-identify, 16% Asian or Pacific Islander, and 8% Hispanic/Latino. Participants were from across the U.S., including Florida (12%), California (5%), Indiana (3%), and other 28 states. Participants were mostly heterosexual or straight (47%), some bisexual (28%), some preferred not to self identify (12%), and homosexual (11%).

We verified a total of 100 participants with 11,062 conversations out of which 1,452 (13.13%) conversations were marked as unsafe by participants. We divided the media files into three categories:

- **Personal Media Files:** These are media files (images, videos and GIFs) that the user sends from their own phone. Such images are personal because they are saved in the user's own mobile gallery.
- **Instagram Media Files:** These are images, profile posts or stories publicly uploaded on Instagram platform whose links are then shared in the conversation by users.
- **Other Media URLs:** These include links to external platforms (e.g., YouTube, Giphy) that users might share in their private conversations.

## 3.5 Analytical Methods

Given the large size and multimodal nature of Instagram data, we leveraged a number of quantitative techniques to answer our overarching research questions.

*3.5.1 Image Analysis Techniques.* A conversation on Instagram usually involves two or more people. To quantify the messages sent, we first divided the data set into media files that were sent by the participants and those they received. Then, we use two techniques to process images and screenshots. We decided to use the MSCOCO deep learning model [55], which, given an image, generates a textual caption. This is a suitable tool to answer RQ2, since we can apply natural language processing techniques to these captions to identify important themes in images shared in safe and unsafe conversations. Upon careful manual inspection, we found that some of the images were screenshots of computer or mobile displays, so we employed an OCR tool to extract the text inside them. Extracting the text in screenshots allowed us to identify important themes discussed in them. Details about the two tools are as follows:

*1) Deep Learning Model to Generate Captions.* To understand the content of these images we used the TensorFlow implementation of the image-to-text model described in [55] to generate captions for all the images. This paper provided a generative model based on computer vision and machine translation. It can be used to generate natural sentences describing an image. For each caption of the image, the tool also produces a confidence score that represents the confidence that the model has in the caption it generated. We manually looked at 20 images and found that captions with confidence score lower than $1 \cdot 10^{-7}$ were not accurate (a total of 5,773 images were removed out of 43,953 images (13.13%), and hence we discarded them from our analysis). Typically images that contain a lot of text or are blurry did not perform well with this model.

*2) Screenshot Detection and Text Extraction.* Some of the images in our dataset were actually screenshots of other conversations or of a phone screen. A screenshot is an image that shows the contents of a computer or mobile display [43]. Screenshots capture what a person might be seeing on their screen so that they can share it with others or save for later (archiving the past). This is helpful for people to show others exactly what they are seeing online – a message, an error on a website, or just a simple meme. Therefore, CHI community [25] has used screenshots analysis to learn more about users' workflow and pain points.

Adolescents may send each other screenshots when they aim to capture something that someone sent directly to them. These screenshots have the potential to involve unsafe content. Receiving an uncomfortable message may prompt an adolescent to screenshot the message and show their friends for advice. Since screenshots are likely to contain a considerable amount of text, the caption extraction tool mentioned before is not appropriate to analyze them. Instead, we opt to develop a tool to automatically identify screenshots and later apply optical character recognition (OCR) techniques to extract the text contained in them for further analysis.

We developed a method to automatically identify screenshots in our dataset using the following features:

- **Dimensions of the image:** Screenshots capture the entire phone display. As a result, the size of a screenshot falls within certain dimensions. For example, the image dimensions of a screenshot taken on a 4.7 inch iPhone 6 are $750 \times 1334$ pixels. By consulting Apple and Android documentation, as well as using manual observation, we identified common form

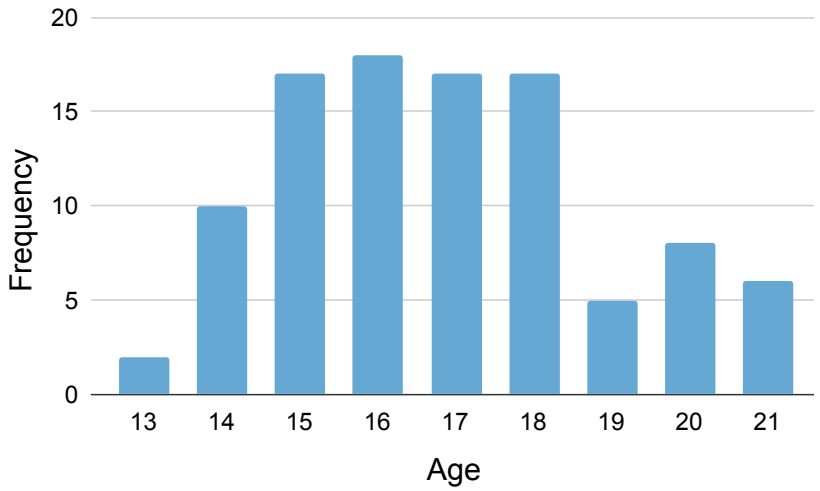


**Figure 2: Frequency of participant ages.**

factors used by mobile phones and identified 60 different image dimensions for screenshots.

- **Presence of a status bar in the image:** We also look for the presence of a status bar in the images. These status bars often contain the time, battery indicator, and other symbols that help detect whether the image is a screenshot.
- **Number of words in the image:** A screenshot can also be cropped distorting the dimensions and possible loss of status bar, therefore classifying these pictures as screenshots solely on their dimensions is not useful. Our assumption was that images that contain a large number of words are very likely to be screenshots. We extracted the text using OpenCV library for python.

After separating the screenshots we pre-processed the text in them by removing empty strings, punctuation, and any non-alpha/numeric characters. As a result, we established a string of words associated with each screenshot.

*3.5.2 Text Analysis Techniques.* After generating the captions for all the images, we separated the screenshots from regular images and extracted the text in them we move on to performing analysis on the extracted text and captions of the images using TF-IDF and the Empath tool [10] – a large-scale language modeling approach. Using Figure 1 as our reference, we divided our analysis into two analytical phases - we first process images and then analyze them. Then provide a high level overview of these techniques.

*1) TF-IDF.* To better understand important themes appearing in the images shared in safe and unsafe conversations, we apply Term Frequency/Inverse Document Frequency (TF-IDF) on the captions and OCR text that we extract. TF-IDF is a product of two metrics, namely Term Frequency (TF) and Inverse Document Frequency (IDF). The idea is that we can infer the words that are most representative of the images and the screenshots shared in both safe and unsafe conversations. In its simplest form, TF is a measure of how frequently a word 't' is found in a caption or screenshot 'd'. IDF is a logarithmic scaling of the fraction of the number of images/screenshots containing word 't'. TF-IDF is then computed by multiplying TF and IDF, as shown in Equation 1.

$$\text{TF-IDF}(t, d, D) = TF(t, d) \times IDF(t, D) \quad (1)$$

Once TF and IDF are obtained, we pick the top keywords. The output of TF-IDF is a weighted metric that ranges between 0 and 1. The closer the weighted value is to 1, the more important the term is in the corpus.

*2) Empath Scoring.* To provide a better quantitative characterization of the captions extracted from images shared in safe and unsafe conversations we leverage Empath [10], a tool for analyzing text across different lexical categories. Empath works by using seed words with member terms which are generated by querying a Vector Space Model (VSM) that was previously trained on a corpus of over 1.8 billion words of fiction. With those seed words, Empath counts category terms in a document related to those seed words and returns the counts for each of its 200 categories. These raw counts can then be normalized over the words in the document to be between 0 and 1.

In this research, we focused on whether images convey negative and positive emotions. To this end, we used the Empath categories that were selected by previous research [44]. From the 194 total Empath categories, they selected the following (a) 15 categories related to hate, violence, discrimination, and negative feelings, and (b) 5 categories related to positive matters in general. In the following, we list all 20 categories:

- **Negative:** aggression, anger, disgust, dominant personality, hate, kill, negative emotion, nervousness, pain, rage, sadness, suffering, swearing terms, terrorism, violence.
- **Positive:** joy, love, optimist, politeness, positive emotion.

*3.5.3 Qualitative Analysis Approach.* In addition to our computational and quantitative analyses of the data, we also iteratively performed a qualitative content analysis [18] on images and screenshots to identify key differences and patterns between the safe versus unsafe conversations. This helped us to uncover nuances in

| Media Category | Safe | Unsafe |
|---|---|---|
| Personal Media Files | 35,815 | 2,305 |
| Instagram Media Files | 96,590 | 2,638 |
| Other Media URLs | 108,755 | 4,413 |
| **Total** | **241,160** | **9,356** |

**Table 1: Number of media files in each category split into safe vs unsafe conversations.**

| Media Category | Total | Participants | Others |
|---|---|---|---|
| Personal Media Files | 38,120 | 14,098 | 24,022 |
| Instagram Media Files | 99,228 | 535 | 98,693 |
| Other Media URLs | 113,168 | 9,867 | 103,301 |
| **Total** | **250,516** | **24,500** | **226,016** |

**Table 2: Number of media files in each category and the difference in stats between media files sent by participants and those they received by others.**

the data and provide exemplar cases to help unpack our results. For instance, the findings that participants often shared screenshots with their friends and that these screenshots often contained risky interactions with others was an emergent qualitative finding from our early qualitative inspection of the data. Based on this insight, we included screenshot detection and text extraction in our analysis pipeline. We also highlight other emergent themes throughout our paper, such as many of the unsafe conversations being sexual in nature. Next, we present our findings based on our research questions.

**Disclaimer.** *The following content contains language and images that some readers might find offensive, sensitive, or triggering. Out of respect to our participants' lived experiences, we do not edit any profanity, sensitive language, or innuendos from quoted excerpts or images, however, triggering images have been blurred.*

## 4 FINDINGS

In this section, we present the findings of our analysis aimed to address our two research questions. First, we analyze how media varies between safe and unsafe conversations (RQ1). Second, we analyze the key characteristics of safe and unsafe conversations (RQ2).

### 4.1 RQ1 –Between-group Differences: Statistical variations between safe and unsafe conversations

In total, participants labeled 11,062 conversations, 1,452 of which were marked as unsafe (13.13%). Our dataset contained a total of 43,953 images extracted from Personal and Instagram Media Files out of which the total number of images in the unsafe conversations is 3,009.

*4.1.1 Types of media files shared in safe and unsafe conversations.* Table 1 provides an overview of the types of media files that appeared in safe and unsafe conversations. As it can be seen from the table, more media files were shared in safe conversations. A $\chi^2$ test gave the result $p < 0.00001$; significant at $p = .01$).

On average the number of media files shared in an unsafe conversation was 6.563 (standard deviation 62.00) and the average number of media files shared in a safe conversation was 22.36 (standard deviation 169.94). This shows that significantly more media files were shared in safe conversations than in unsafe conversations.

*4.1.2 Media sharing characteristics of participants versus others.* We give an overview of the total number of media files shared by participants and by others in Table 2. As it can be seen, participants received more media files overall (226,016 out of 250,516 i.e. 90.22%) than they sent (24,500 out of 250,516 i.e. 9.78%). A reason for this disparity might be that 1,288 (11.64%) of the conversations were between a participant and multiple people. We then focused on the sharing activity of participants and others in safe and unsafe conversations. We started by looking at the entirety of conversations (including text messages and media), and then analyzed the sharing of media only. Safe conversations are usually longer, with both the participant and others going back and forth (mean length 159.33, median length 6.0, standard deviation 1011.05), while unsafe conversations are shorter (mean length 83.93, median length 3.0, standard deviation 588.69).

Figure 3(a) shows the cumulative distribution function (CDF) of the average number of messages sent by participants and others in safe conversations. A CDF shows the proportion of values less than or equal to X at a certain point. CDF plots are useful for comparing the distribution of different sets of data. For example in Figure 3(a) participants sent more messages in conversations on average, with 50% of users posting more than 147 messages in an average conversation. To assess if the difference between these distributions is statistically significant, we ran a two sample Kolmogorov-Smirnov test (KS test) [31]. We found that the differences between the distributions were statistically significant at $p = 0.01$ ($D = 0.386, p = 0.0026$). This means that the number of messages received by participants was significantly greater than the number of messages sent by them. We also compared the number of media files that the participants and others shared in safe conversations in Figure 3(b). Unlike text messages, participants send on average less media than others. This is confirmed by another KS test, which shows that the two distributions have a statistically significant difference at $p = 0.01$ ($D = 0.432, p = 0.00046$).

Next, we turn our attention to unsafe conversations. The total number of media files sent in unsafe conversations is 9,356 and out of those 371 were sent by participants, while 8,985 media files were received. Figure 4(a) shows the CDF of the average number of messages sent by participants and by others in an unsafe conversation. We saw that others send more messages in unsafe conversations (by 87.11% as compared to in safe conversations). We confirmed this using KS test that returned the results were statistically significant at $p = 0.01$ ($D = 0.115, p = 9.492 \cdot e^{-60}$).

We then look at media shared in unsafe conversations. Figure 4(b) shows the CDF of the number of media files shared by participants and others. In unsafe conversations, participants were more likely to receive media files than send them; 85% of the participants did not send any images at all in the unsafe conversations. A KS test found

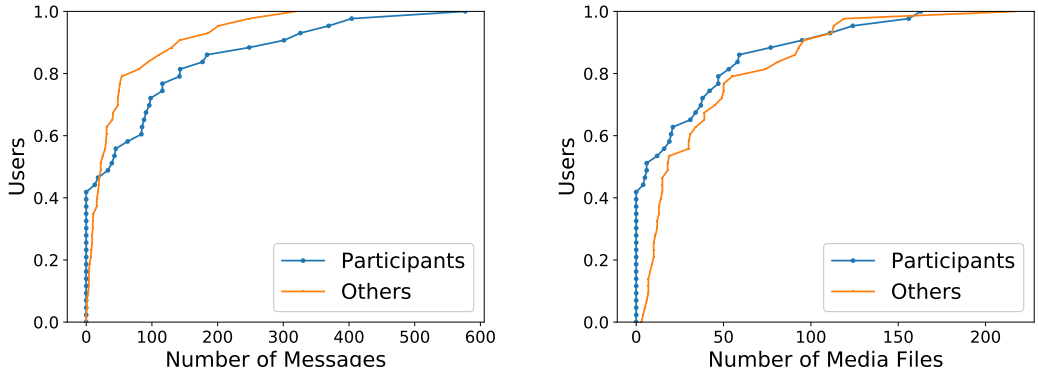


**Figure 3: CDF of (a) the average number of messages and (b) the average number of media files shared by participants vs others in safe conversations.**

that the difference in distribution here is statistically significant at $p = 0.01$ ($D = 0.102, p = 6.589 \cdot e^{-25}$).

To better understand the reason behind this discrepancy we further analyzed the message exchange activity between participants and others in unsafe conversations. For 87.2% of unsafe conversations the participant never replied to the other party, while in the remaining conversations, 13.56% of users stopped responding after the first unsafe image was sent. This suggests that youth tends to disengage from conversations that they consider unsafe, and might explain the reason why the number of media files sent by others is higher than those sent by participants in the unsafe conversations.

## 4.2 RQ2 –Image Characteristics: Key Differences between images shared in safe/unsafe conversations

In this section, we analyze the characteristics of images shared in safe and unsafe conversations. We first separated images from other media files (media files include images, videos and GIFs). The total number of media files in our dataset was 250,516 including personal media files, Instagram media files and others. We then found the total number of images shared in DMs which was 43,593 out of the total 250,516 (17.4%). When qualitatively reviewing the images in our dataset, we noticed that some of them were screenshots of other conversations that participants were either receiving or sharing with their contacts. After segregating the screenshots, we identified a total of 4,926 screenshots out of a total of 43,593 images (11.21%) in our dataset. We then analyzed the regular images and the screenshots separately.

*4.2.1 Characterizing Regular Images in Safe and Unsafe Conversations.* As introduced in 3.5, we used the MSCOCO Image Captioning Model [55] to automatically characterize the subjects and content of regular images in our dataset. We then calculated the Term Frequency-Inverse Document Frequency (TF-IDF) of each word in the generated captions to identify the most representative keywords of the set of images that appear in safe and unsafe conversations.

The results are a set of words that best characterize the safe and unsafe images, compared to all images in our dataset. Table 3 shows the top keywords identified for images shared in safe and unsafe conversations. The unsafe category contains keywords that are indicative of people ("girl," "man," "person," "woman"), while the safe category contains keywords that are indicative of objects (e.g., "book," "shoes," "pizza"). The table also shows that the TF-IDF for the words in unsafe conversations is higher than the words in safe conversations. Some of the captions generated for the images in safe conversations were:

> "a book sitting on top of a wooden desk."
> "a bunch of different colored ties hanging from a wall."
> "a bunch of items that are on a table."

Whereas the captions generated for some of the images in unsafe conversations were:

> "two girls sitting on grass."
> "a man in a suit and tie holding a toothbrush."
> "a woman holding a teddy bear in her arms."

We also analyzed whether images convey negative or positive emotions. To perform emotional characterization of the captions extracted from images shared in safe and unsafe conversations we leveraged Empath [44]. Figure 5 shows a heat map of the average normalized Empath scores for both safe and unsafe images across the chosen lexical categories. We can see from the chart that unsafe conversations were more emotionally charged than safe conversations. Unsafe conversations averaged higher scores in every category for both positive and negative categories.

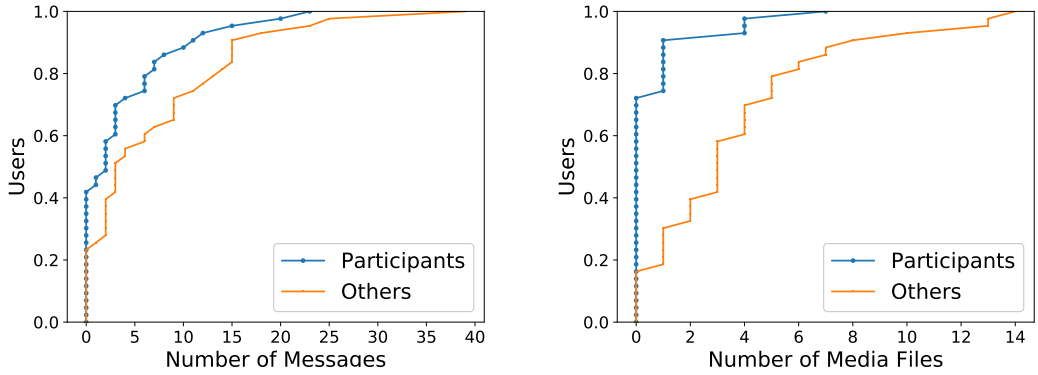


**Figure 4: CDF of (a) the average number of messages and (b) the average number of media files shared by participants vs others in unsafe conversations.**

| Safe Common Words | TF-IDF | Unsafe Common Words | TF-IDF |
|---|---|---|---|
| near | 0.0255 | girl | 0.4512 |
| skyline | 0.0198 | hand | 0.4373 |
| cell phone | 0.0197 | boy | 0.4212 |
| hydrant | 0.0156 | teddy | 0.4100 |
| surface | 0.0140 | bear | 0.4095 |
| rain | 0.0129 | woman | 0.4057 |
| pizza | 0.0134 | black | 0.4042 |
| plane | 0.0112 | white | 0.3001 |
| coffee | 0.0108 | suit | 0.2980 |
| pug | 0.0078 | man | 0.2500 |

**Table 3: List of top 10 keywords from captions generated for Safe Images using TF IDF (on the left) and list of top 10 keywords from captions generated for Unsafe Images using TF-IDF(on the right).**

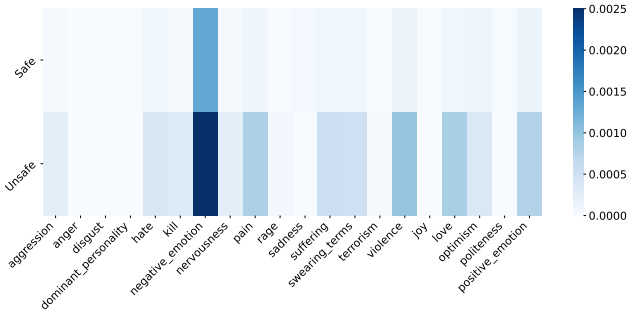


**Figure 5: Average Empath values for all images in safe and unsafe categories.**


**Figure 6: Example of images shared in an unsafe group conversation. The image on the left was sent by a girl when asked for her phone number in a group chat. The image on the right was then sent by a guy implying self-harm. Here, the participant did not send any image/message but was made to feel uncomfortable by her group mates. Note: These images have been recreated to ensure the confidentiality of our participants.**


**Figure 7: Example of memes shared in unsafe conversations. The image on the left is a hateful meme linking gender transitions to suicide. The image on the right was a part of another unsafe conversation where a guy is asking a female participant for nudes. The images are part of different conversations.**

Next, to better understand the type of images shared in safe and unsafe conversations, we manually inspected the images in each category. We found that images shared in unsafe conversations often contained more provocative pictures for example photos of semi-naked women. We also observed instances of images containing insulting or offensive hand gestures shared in unsafe conversations, as well as memes with sexual innuendos. Another theme that we observed in unsafe conversations was selfies of girls or women (not

sexual) being shared by others with participants, with them commenting about their appearance. Cyberbullying or being offensive was a highlight of these conversations, for instance, in an unsafe conversation a guy shared a selfie of another girl doing the middle finger gesture and said that she is a "whore." Figure 6 shows an example of an exchange of images in an unsafe conversation.

Following the findings made through Empath, our manual inspection found that unsafe images were mostly of people being angry, violent, or being involved emotionally. We also found that unsafe images sometimes depicted dark or sarcastic humor usually revolving around death or violence. Figure 7 shows some examples of memes shared in unsafe conversations.

*4.2.2 Characterizing the Use of Screenshots.* As mentioned previously, our dataset contains 4,926 screenshots in total. In the 1,452 unsafe conversations in our dataset, we found 386 screenshots. We quantitatively characterize the content contained in screenshots, focusing on differences between those posted in safe and unsafe conversations. To this end, we extracted the top keywords contained in screenshots shared in safe and unsafe conversations by using TF-IDF. Table 4 reports the top keywords contained in screenshots in safe and unsafe conversations. For our analysis, Term-frequency is "the number of times a word occurs in a screenshot" and Document-frequency is "the number of screenshots a word appears in". Next, we provide more details about unsafe and safe screenshots and go over some representative examples of the screenshots shared in unsafe and safe conversations.

*Unsafe Screenshots: Sexual Solicitations and Harassment with Acquaintances/Friends.* As mentioned when introducing the dataset, we asked participants to further label unsafe conversations with information on the relationship that they have with the other party in the conversation (e.g., family, significant other, friend, acquaintance, stranger, other). Most (45%) of the conversations containing screenshots were with acquaintances, 20% with friends, 5% with an ex, and for the remaining 30% participants chose not to label the relationship. Looking at this information in relation to unsafe conversations where screenshots were shared helped us understand the context in which young adults shared screenshots with each other, and the reasons why they were sharing them. In our dataset, we only found one unsafe conversation where the other party was labeled as a stranger by the participant. This might indicate that youth tend to share screenshots of their online activities and experiences with people who they know and consider screenshots sent by them more seriously that could cause them feel unsafe.

The top words in screenshots shared in unsafe conversations contain words that are sexual, for example *fuck, dick, love*. Although these words are not always presenting unsafe interactions and can be used in safe conversations regarding sexual health, they were present in many unsafe conversations flagged by participants in our study. In addition, TF-ID results demonstrate that usually these screenshots include negative emotions indicating distressful situations such as *disgusting, frightening, harmful, freak, nasty*. Screenshots that were shared in unsafe conversations were usually sent in the context of showing sexual interest in the other person including sexual solicitations or requesting nudity or porn. For example, a guy shared a screenshot with a girl (one of the participants), which included sexual harassment:

| Safe Common Words | TF-IDF | Unsafe Common Words | TF-IDF |
|---|---|---|---|
| load | 0.04190 | man | 0.44510 |
| identical | 0.03314 | bra | 0.34370 |
| host | 0.03111 | fuck | 0.24001 |
| hour | 0.02456 | condoms | 0.23998 |
| hulu | 0.02440 | disgusting | 0.23409 |
| humanity | 0.02335 | frightening | 0.23305 |
| hung | 0.02334 | service | 0.23140 |
| hungry | 0.02332 | harmful | 0.23000 |
| icon | 0.02308 | head | 0.22980 |
| ideas | 0.02248 | health | 0.22150 |
| horrible | 0.02185 | age | 0.14251 |
| ignorance | 0.02098 | feminazi | 0.14037 |
| ignore | 0.01199 | interact | 0.14001 |
| image | 0.01166 | invite | 0.13003 |
| immediately | 0.01140 | joint | 0.04609 |
| immersive | 0.01129 | freak | 0.04025 |
| impact | 0.01114 | boys | 0.04014 |
| impetuous | 0.01112 | mouthpiece | 0.03000 |
| importance | 0.01108 | nasty | 0.02098 |
| impress | 0.00278 | bitch | 0.02020 |
| improving | 0.00245 | love | 0.01451 |
| identical | 0.00198 | nicotine | 0.01237 |
| inappropriate | 0.00035 | hoe | 0.01142 |
| hose | 0.00004 | banana | 0.00410 |
| incident | 0.00002 | dick | 0.00209 |

**Table 4: List of top 25 keywords from text extracted from screenshots shared in Safe Conversations using TF IDF (on the left) and list of top 25 keywords from text extracted from screenshots shared in Safe Conversations using TF-IDF(on the right).**

> "*...want to stick my dick in you and destroy you but also want to treat you like a fucking princess...*"

There were also many instances of harassment, including making fun of sexual orientation or identity, religion, or points of view. For instance, in a group conversation, the people that the participant thought were his friends, grew against him and harassed him over his sexual orientation and identity and made fun of him. They send screenshots of insulting jokes about "femboys", males whose appearance and behavioral traits are regarded as conventionally feminine, which made the participant feel unsafe and insulted.

> "*i dont have a thong yet. idk why it is in the gay furry meme insta community if u have panties u think ur god.femboys are dumb.one was like complaining how they went to victorias secret snd get looked at weird when they went to try it on*"

Figure 8 shows some more examples of screenshots shared in unsafe conversations.

*Safe Screenshots: Asking for Advice Regarding Difficult Situations.* The top words in screenshots shared in safe conversations are more varied in nature. Many of the screenshots that were shared in safe conversations were related to the participant or the other person asking for advice on how to answer or deal with an uncomfortable situation. They shared a screenshot and asked the other person for

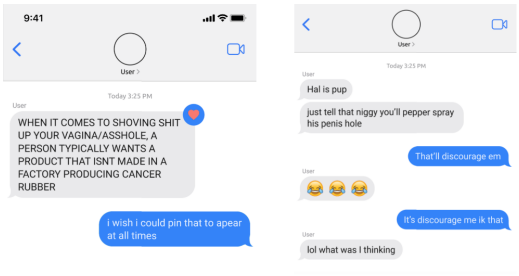


**Figure 8: Examples of screenshots shared in unsafe conversations. The screenshot on the left was sent in an unsafe conversation where the other person was saying that he is against the use of contraceptives and that the person who he is conversing with also agrees, so the participant should agree too. This conversation also included messages about sex toys. The screenshot on the right was sent by a friend who is sharing the suggestion she gave to another friend on how to deal with an ex. It also includes an "invisible" media. Such images can be viewed only once. The sender usually does not want the receiver to be able to 'save' such images but only view them.**

support. Relevant TF-IDF words related to such instances include *horrible, ignorance, inappropriate, incident*. For example, a participant shared a screenshot of a conversation he had with a girl with his friend. In the screenshot, the girl threatens him that she knows his parents and that he must stop texting her or her friends. The participant answers that he is just being nice and not creepy. In the conversation where this screenshot is shared, the participant's friend supports him and says that the girl is just mean and he should drop all communication with her, before things escalate.

This example shows that youth share screenshots of their conversations with others to their friends to seek advice. The participant sent the following screenshot (excerpt) to one of his friends asking for advice.

> **Girl:** "*I know your parents*"
> **Participant:** "*What did it*"
> **Participant:** "*What did i do*"
> **Girl:** "*Just don't text my friends or me ever again before you regret it creep*"
> **Participant:** "*What did I even do*"
> **Participant:** "*I'm not even doing anything creepy*"
> **Girl:** "*I'm just trying to be nice*"

Interestingly, this example suggests that the participant's previous interactions with the girl may have made her feel unsafe, but in turn, he sought advice from a friend because her accusations made him feel uncomfortable and somewhat threatened as well. This interaction illustrates how youth have meta-conversations with one another to sort out the complexities of experiences that made them or someone else feel unsafe. In another instance, a participant shared a screenshot of a conversation in which she was asked to send nude pictures of herself by a stranger. In the screenshot the stranger calls the girl his "fuck buddy." She shared this screenshot in a conversation with a friend, who advised her to ignore "creepy" people/messages and it is probably a scam. The following example is an excerpt from another screenshot showing that youth share screenshots of unsafe conversations with their friends to seek support and the best way forward.

> "*You're my...Fuck buddy. Screenshot this and send me your answer and I'll send you the vid*"

Overall, many screenshots that were shared in unsafe conversations were between friends/acquaintances contained sexual and offensive content that made the youth feel uncomfortable or unsafe. While in safe conversations, youth shared screenshots of unsafe interactions and asked their friends how to handle these situations.

## 5 DISCUSSION

Our results provided many valuable insights about safe and unsafe private conversations of teens and young adults online. Corresponding to RQ1, we found that unsafe conversations are more likely to contain media images compared to safe ones. We also found that participants tend to send fewer messages and media when engaged in unsafe conversations, and that most media is actually sent by the other party. When investigating RQ2, we found that regular images shared in unsafe conversations are more likely to contain people, while those shared in safe conversations are more likely to contain objects. When analyzing screenshots, we also discovered that participants often shared screenshots of other conversations with their friends, including screenshots of unsafe conversations that made them uncomfortable, to perhaps seek support or ask for advice. Below, we discuss the human-centered insights gained regarding youth risk behavior, as well as the implications for future work towards building machine learning techniques for the automatic detection of such risks.

### 5.1 Understanding What Makes Youth Feel Safe versus Unsafe Online

Our research is one of the first to examine the characteristics of private message conversations that make youth feel uncomfortable or unsafe on social media. A decent percentage (13.12%) of the private message conversations shared by our participants made them or someone else feel unsafe or uncomfortable. Further, privately shared personal images were more indicative of risk behavior than publicly media re-shared privately. These findings confirms that many of the risks youth encounter online occur in private spaces and validates Pinter et al.'s earlier recommendation that researchers need to move beyond youth self-reports and publicly scraped social media posts when gaining insights into youth online risk behavior [45]. A key implication is that researchers examining the online risk behavior of youth must continue to take the Herculean efforts needed to collect ecologically valid datasets, including private and intimate social media interactions, so that their research has real-world impact on the online safety and protection of youth. Yet, such efforts require great care to protect youth from both the ethical and legal ramifications of such research [48].

Second, we uncovered valuable and new knowledge about youth and their experiences that made them feel unsafe online. For instance, we quantitatively validated that youth send significantly fewer messages in conversations where they felt unsafe or uncomfortable. Qualitatively, we observed that once youth perceived an

interaction as unsafe, they often quickly disengaged by not responding. This finding is confirmatory evidence to Jia et al.'s claim that youth may take protective actions once their safety concerns have been heightened [23], rather than proactively protecting themselves through making fewer online disclosures (i.e., being more private) or not interacting online with people who are strangers. To add to the finding that adults shared a wider array of images than youth [22], we found that youth shared a wider array of safe images than unsafe images, which were mostly risque photos.

Contrary to the assumption that youth are unable to accurately assess and effectively cope with online risks, our findings suggest that the youth know when to disengage from unsafe interactions and they actively seek support through the sharing of these experiences with others. In some cases, participants shared screenshots to disclose instances of online abuse (e.g., unwanted sexual advances or harassment) with their friends to garner support. In other cases, we even saw evidence of perpetrators sharing the repercussions of their mistakes (e.g., backlash from a victim) with friends to get advice on how to make amends or avoid getting in trouble. These findings are reminiscent of Razi et al.'s [46] work that found that adolescents seek advice and support from strangers about their online sexual risk experiences. However, our study is the first to shed light on the meta-level discourse youth have with their friends regarding their online risk experiences, in addition to the actual risks they experience in private online spaces. Thus, future work should consider how youth not only experience online risks for the purpose of developing effective interventions but also how youth disclose these experiences to others to process, heal, and/or learn from these negative experiences.

## 5.2 Implications for Automated Risk Detection Systems for Youth

Our analysis provides an important first step to identify trends in the media shared in unsafe conversations of youth to enable automated (machine learning based) detection of online risk behavior going forward. In particular, we established the distinct ways youth shared media, whether those they share or receive, are distinct across private conversations they perceived to be safe or unsafe. The characteristics of the media we identified in this work (e.g., nature and volume of social engagement involving media, sharing of screenshots, and so on) could serve as features in a supervised machine learning setup to help train models that detect risk. In contrast to existing risk detection systems, such models are likely to benefit on two fronts: first, they will be able to harness the perspectives of the youth victimized in unsafe conversations as sources of ground truth data; and second, they will be able to draw, in a grounded way, from the insights and interpretations we gathered about media sharing behaviors. As Kim et al [27] argued in their work, these human-centered approaches to detecting online risk are not likely to be more realistic of the actual risk experienced by youth online, but are likely to be more translatable in the real world given their rich ecological validity. Moreover, since the conversations we studied are actually annotated by the youth who have themselves felt uncomfortable, we can extrapolate the key points from this research to automated risk detection in other conversations that show similar traits.

This study also underscores the value of a multimodal approach to online risk detection. The vast majority of risk detection algorithms, such as those detecting cyberbullying or harassment use textual features [35, 51]. Our analysis of the subjects and content of the images shared and received by youth indicate that risk detection algorithms need to include features extracted from images as well, thereby adopting a more comprehensive approach to capturing the multimodal nature of unsafe conversations. With platforms like Instagram – the platform in consideration – getting increasingly popular among youth, such a multimodal approach will only accrue greater significance in risk detection systems.

## 5.3 Implications for Design

There are several implications for the design of new safety features for social media that can be derived from our findings. First, it may be useful for private messaging platforms to identify whether youth users on their platforms are engaging with friends, acquaintances, or strangers. We noticed several instances where unsafe conversations did not appear to be with someone our participants had a close relationship with. Further, screenshots of unsafe interactions were more often shared within trusted relationships in order to seek advice and support. Indeed, Instagram recently implemented a major shift in policy, where adult users are not allowed to private message with minors who they do not follow [20]. Another design recommendation would be for social media platforms to assess whether privately shared images being sent to minors are unsafe, then make recommendations to young users on how to handle such situations. One common approach used by our participants, for instance, was to simply disengage by not responding to an offensive message. These recommendations could come in the form of nudges, similar to those suggested by Masaki et al.'s work [37] for helping adolescents avoid privacy and safety threats. While their work suggested the use of negatively framed social nudges (e.g., "90% of users would not share a photo without permission") to reduce youths' own inappropriate media sharing, our findings suggest that encouraging youth to ask friends for help in the event that they receive an unsafe photo may potentially be an even more effective nudge, helping them garner needed advice and support. Such a nudge could be designed to allow the user to easily capture a screenshot and share an unsafe interaction with a trusted friend. Yet, many of these design recommendations rely on the proactive and accurate detection of potentially unsafe situations, so that context-appropriate nudges can be employed.

## 5.4 Limitations and Future Work

Due to the sensitive nature of our study, we had to make several trade-offs in the design of our study that were necessary to protect our participants. First, while we analyzed a large amount of Instagram data, our participant sample size was relatively small due to the technical, practical, and ethical complexities of collecting private conversations from adolescents and young adults. Yet, we have very rich data spanning thousands of media per participant, and our results on these 250,516 media files did enable us to identify statistically significant differences between media shared in safe versus unsafe conversations. Further, there may be instances where unsafe media was removed from the data prior to upload

or otherwise unavailable for analysis. For legal reasons, we asked youth to remove any instances of child pornography from their data; therefore, we were not able to (nor did we want to) capture or analyze this type of high-risk and illegal media. We can infer that there were likely riskier images shared among youth, even though what they did share was still quite risque. Additionally, when processing old image links from historical Instagram data, many of the media were no longer available.

From a technical standpoint, our Caption Generating Model [55] generated high-quality captions for only high-resolution images. In the cases where images were blurry and the confidence-level in the caption was low, we did not include these images in our analysis. A future research direction is to develop better deep learning based systems that are able not only to recognize the subject of an image, but also to characterize their context, with the goal of measuring the inherent risk of images other than their content. Further, we used OpenCV to extract words/text contained in the images, which has demonstrated inferior performance to cloud-based systems, such as Google Vision API [13]. However, we felt that this trade-off was necessary to protect the sensitive data entrusted to us by our participants and encourage researchers to also consider making such choices when they encounter similar cases to ours.

In our research, we used risk-flagged data from the participants themselves. However, it is known that youth may under-estimate risks encountered online [60]; therefore, we found some instances of unsafe conversations (e.g., the sharing of a pornographic video) that were not flagged as unsafe. Therefore, while we can say with confidence that risk-flagged conversations contained unsafe interactions, it may be that non-flagged conversations may have also contained risk. Finally, there are several ways in which future research can build upon our study. For example, we only made a distinction between safe and unsafe conversations. Future research could distinguish between different types of risks (e.g., sexual predation, cyberbullying, mental health) as well as whether the individual was the victim or perpetrator of that risk to determine if the same patterns hold as in our results. In our future work, our goal is to employ both qualitative methods to deeply understand youth risk behavior and machine learning approaches to more accurately detect the various types of risks youth encounter online.

## 6 CONCLUSION

In this paper, we presented a data-driven study of images shared in safe versus unsafe DM conversations of youth on Instagram. We found that youth share different types of media in direct messages on Instagram. They not only share private images of themselves but they also share public links of other people's images that are present on Instagram. Our findings suggest that youth stop engaging in predatory conversations and that unsafe conversations are shorter, one sided and lesser media files are shared in them. Interestingly we found that people share screenshot images in conversations with their friends even if that conversation later on makes them uncomfortable. Our work takes the necessary steps in identifying key characteristics in safe and unsafe conversations that can be used to identify risks.

## ACKNOWLEDGMENTS

This research is supported in part by the U.S. National Science Foundation under grants #IIP-1827700, #IIS-1844881, 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. We'd like to thank Tess Conroy for helping design the experiments for separating screenshots from images, and Luke Shirley for recreating the screenshots. Also, special thanks to all the participants who donated their data.

## REFERENCES

[1] Monica Anderson and Jingjing Jiang. 2018. Teens, Social Media & Technology 2018 | Pew Research Center. https://www.pewresearch.org/internet/2018/05/31/teens-social-media-technology-2018/
[2] 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/
[3] 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. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW3, Article 231 (Jan. 2021), 46 pages. https://doi.org/10.1145/3432930
[4] Susan B Barnes. 2006. A privacy paradox: Social networking in the United States.
[5] U.S. Census Bureau. 2021. U.S. Census Bureau QuickFacts: United States. https://www.census.gov/quickfacts/fact/table/US/PST045219
[6] d. boyd. 2007. Why Youth (Heart) Social Network Sites: The Role of Networked Publics in Teenage Social Life.
[7] danah boyd. 2014. It's Complicated: The Social Lives of Networked Teens. http://www.jstor.org/stable/j.ctt5vm5gk
[8] Suzan M Doornwaard, Megan A Moreno, Regina JJM van den Eijnden, Ine Vanwesenbeeck, and Tom FM Ter Bogt. 2014. Young adolescents' sexual and romantic reference displays on Facebook. *Journal of Adolescent Health* 55, 4 (2014), 535–541.
[9] Rebecca Dredge, John Gleeson, and Xochitl De la Piedad Garcia. 2014. Cyberbullying in social networking sites: An adolescent victim's perspective. *Computers in human behavior* 36 (2014), 13–20.
[10] Ethan Fast, Binbin Chen, and Michael S. Bernstein. 2016. *Empath: Understanding Topic Signals in Large-Scale Text.* Association for Computing Machinery, New York, NY, USA, 4647–4657. https://doi.org/10.1145/2858036.2858535
[11] Robert J. Fisher. 1993. Social desirability bias and the validity of indirect questioning. *Journal of consumer research* 20, 2 (1993), 303–315.
[12] 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/
[13] Google. 2021. Vision AI | Derive Image Insights via ML | Cloud Vision API. https://cloud.google.com/vision
[14] Heidi Hartikainen, Afsaneh Razi, and Pamela Wisniewski. 2021. *'If You Care About Me, You'll Send Me a Pic' - Examining the Role of Peer Pressure in Adolescent Sexting.* Association for Computing Machinery, New York, NY, USA, 67–71. https://doi.org/10.1145/3462204.3481739
[15] Heidi Hartikainen, Afsaneh Razi, and Pamela Wisniewski. 2021. Safe Sexting: The Advice and Support Adolescents Receive from Peers regarding Online Sexual Risks. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–31.
[16] 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. , 250–259 pages.
[17] Homa Hosseinmardi, Sabrina Arredondo Mattson, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra. 2015. Detection of Cyberbullying Incidents on the Instagram Social Network. arXiv:1503.03909 http://arxiv.org/abs/1503.03909
[18] Hsiu-Fang Hsieh and Sarah E Shannon. 2005. Three approaches to qualitative content analysis. *Qualitative health research* 15, 9 (2005), 1277–1288.
[19] Yuheng Hu, Lydia Manikonda, and Subbarao Kambhampati. 2014. What We Instagram: A First Analysis of Instagram Photo Content and User Types. https://ojs.aaai.org/index.php/ICWSM/article/view/14578
[20] Instagram. 2021. Continuing to Make Instagram Safer for the Youngest Members of Our Community. https://about.instagram.com/blog/announcements/continuing-to-make-instagram-safer-for-the-youngest-members-of-our-community
[21] Mansoor Iqbal. 2021. Instagram Revenue and Usage Statistics (2021) - Business of Apps. https://www.businessofapps.com/data/instagram-statistics/

[22] Jin Jang, Kyungsik Han, Dongwon Lee, Haiyan Jia, and Patrick Shih. 2016. Teens Engage More with Fewer Photos: Temporal and Comparative Analysis on Behaviors in Instagram. https://doi.org/10.1145/2914586.2914602

[23] Haiyan Jia, Pamela Wisniewski, Heng Xu, Mary Beth Rosson, and John Carroll. 2014. Risk-taking as a Learning Process for Shaping Teen's Online Information Privacy Behaviors. *Proc. Computer-Supported Cooperative Work and Social Computing* 2015 (01 2014). https://doi.org/10.1145/2675133.2675287

[24] Lisa M Jones, Kimberly J Mitchell, and David Finkelhor. 2012. Trends in youth internet victimization: Findings from three youth internet safety surveys 2000–2010. *Journal of adolescent Health* 50, 2 (2012), 179–186.

[25] Amy K Karlson, Shamsi T Iqbal, Brian Meyers, Gonzalo Ramos, Kathy Lee, and John C Tang. 2010. Mobile taskflow in context: a screenshot study of smartphone usage. , 2009–2018 pages.

[26] Jung-Eun Kim, Emily C. Weinstein, and Robert L. Selman. 2017. Romantic Relationship Advice From Anonymous Online Helpers: The Peer Support Adolescents Exchange. *Youth & Society* 49, 3 (April 2017), 369–392. https://doi.org/10.1177/0044118X15604849

[27] 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.

[28] Seunghyun Kim, Afsaneh Razi, Gianluca Stringhini, Pamela J. Wisniewski, and Munmun De Choudhury. 2021. A Human-Centered Systematic Literature Review of Cyberbullying Detection Algorithms. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 325 (oct 2021), 34 pages. https://doi.org/10.1145/3476066

[29] J. Kofoed and Malene Charlotte Larsen. 2016. A snap of intimacy: Photo-sharing practices among young people on social media.

[30] Kay l. O'Halloran, Alvin Chua, and Alexey Podlasov. 2014. 25. The role of images in social media analytics: A multimodal digital humanities approach. , 565–588 pages. https://doi.org/10.1515/9783110255492.565

[31] Bernard Lindgren. 2017. Statistical theory.

[32] Ping Liu, Joshua Guberman, Libby Hemphill, and Aron Culotta. 2018. Forecasting the Presence and Intensity of Hostility on Instagram Using Linguistic and Social Features. https://ojs.aaai.org/index.php/ICWSM/article/view/15022

[33] Sonia Livingstone, Leslie Haddon, Anke Görzig, and Kjartan Ólafsson. 2010. Risks and safety on the internet: the perspective of European children: key findings from the EU Kids Online survey of 9-16 year olds and their parents in 25 countries.

[34] Xiao Ma, Jeff Hancock, and Mor Naaman. 2016. Anonymity, intimacy and self-disclosure in social media. , 3857–3869 pages.

[35] Tolba Marwa, Ouadfel Salima, and Meshoul Souham. 2018. Deep learning for online harassment detection in tweets. https://doi.org/10.1109/PAIS.2018.8598530

[36] Alice E Marwick and danah boyd. 2014. Networked privacy: How teenagers negotiate context in social media. *New media & society* 16, 7 (2014), 1051–1067.

[37] Hiroaki Masaki, Kengo Shibata, Shui Hoshino, Takahiro Ishihama, Nagayuki Saito, and Koji Yatani. 2020. Exploring Nudge Designs to Help Adolescent SNS Users Avoid Privacy and Safety Threats. , 11 pages. https://doi.org/10.1145/3313831.3376666

[38] Bridget Christine McHugh, Pamela Wisniewski, Mary Beth Rosson, and John M Carroll. 2018. When social media traumatizes teens.

[39] Jerold D Miller and Shirley M Hufstedler. 2009. Cyberbullying Knows No Borders.

[40] Pushkar Mishra, Helen Yannakoudakis, and Ekaterina Shutova. 2019. Tackling online abuse: A survey of automated abuse detection methods.

[41] Maryam Mohsin. 2021. 10 Instagram Statistics You Need to Know in 2021 [New Data]. https://www.oberlo.com/blog/instagram-stats-every-marketer-should-know

[42] R Morris, Daniel McDuff, and R Calvo. 2014. Crowdsourcing techniques for affective computing. , 384–394 pages.

[43] Lauren North. 2021. What is a screenshot? | TechSmith. https://www.techsmith.com/blog/screenshot/

[44] Raphael Ottoni, Evandro Cunha, Gabriel Magno, P. Bernardina, W. Meira, and Virgílio A. F. Almeida. 2018. Analyzing Right-wing YouTube Channels: Hate, Violence and Discrimination.

[45] 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* (Stanford, California, USA) *(IDC '17)*. Association for Computing Machinery, New York, NY, USA, 352–357. https://doi.org/10.1145/3078072.3079722

[46] 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* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376400

[47] Afsaneh Razi, Seunghyun Kim, Ashwaq Alsoubai, Gianluca Stringhini, Thamar Solorio, Munmun De Choudhury, and Pamela J. 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

[48] Afsaneh Razi, Seunghyun Kim, Munmun De Choudhury, and Pamela Wisniewski. 2019. Ethical considerations for adolescent online risk detection AI systems.

[49] Hugo Rosa, Nádia Pereira, Ricardo Ribeiro, Paula Costa Ferreira, João Paulo Carvalho, Sofia Oliveira, Luísa Coheur, Paula Paulino, AM Veiga Simão, and Isabel Trancoso. 2019. Automatic cyberbullying detection: A systematic review. *Computers in Human Behavior* 93 (2019), 333–345.

[50] Semiu Salawu, Yulan He, and Joanna Lumsden. 2017. Approaches to automated detection of cyberbullying: A survey. *IEEE Transactions on Affective Computing* 11, 1 (2017), 3–24.

[51] Shruthi and Prof Mangala C. 2017. A Framework for Automatic Detection and Prevention of Cyberbullying in Social Media. *International Journal of Innovative Research in Computer and Communication Engineering* 5, 6 (2017), 86–90. www.ijircce.com

[52] 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

[53] H Tankovska. 2021. Instagram: age distribution of global audiences 2021 | Statista. https://www.statista.com/statistics/325587/instagram-global-age-group/

[54] Muhammad Uzair Tariq, Afsaneh Razi, Karla Badillo-Urquiola, and Pamela Wisniewski. 2019. A Review of the Gaps and Opportunities of Nudity and Skin Detection Algorithmic Research for the Purpose of Combating Adolescent Sexting Behaviors. , 90–108 pages.

[55] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2017. Show and Tell: Lessons Learned from the 2015 MSCOCO Image Captioning Challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39, 4 (2017), 652–663. https://doi.org/10.1109/TPAMI.2016.2587640

[56] Yilin Wang and Baoxin Li. 2015. Sentiment Analysis for Social Media Images. , 1584-1591 pages. https://doi.org/10.1109/ICDMW.2015.142

[57] Andy Williams. 2015. Child sexual victimisation: Ethnographic stories of stranger and acquaintance grooming. *Journal of Sexual Aggression* 21, 1 (2015), 28–42.

[58] Pamela Wisniewski. 2018. The Privacy Paradox of Adolescent Online Safety: A Matter of Risk Prevention or Risk Resilience? *IEEE Security Privacy* 16, 2 (2018), 86–90. https://doi.org/10.1109/MSP.2018.1870874

[59] Pamela Wisniewski, Haiyan Jia, Heng Xu, Mary Beth Rosson, and John M. Carroll. 2015. "Preventative" vs. "Reactive": How Parental Mediation Influences Teens' Social Media Privacy Behaviors. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (Vancouver, BC, Canada) *(CSCW '15)*. Association for Computing Machinery, New York, NY, USA, 302–316. https://doi.org/10.1145/2675133.2675293

[60] Pamela Wisniewski, Heng Xu, Mary Beth Rosson, and John M. Carroll. 2017. Parents Just Don't Understand: Why Teens Don't Talk to Parents about Their Online Risk Experiences. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) *(CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 523–540. https://doi.org/10.1145/2998181.2998236

[61] 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. , 3919–3930 pages.

[62] Michele Zappavigna. 2016. Social media photography: construing subjectivity in Instagram images. *Visual Communication* 15, 3 (2016), 271–292. https://doi.org/10.1177/1470357216643220 arXiv:https://doi.org/10.1177/1470357216643220