2022

A Human-Centered Approach to Improving Adolescent Online Sexual Risk Detection Algorithms

Afsaneh Razi
University of Central Florida

A HUMAN-CENTERED APPROACH TO IMPROVING ADOLESCENT ONLINE SEXUAL RISK DETECTION ALGORITHMS

by

AFSANEH RAZI
B.S. University of Tehran, 2013
M.S. Islamic Azad University Central Tehran Branch, 2015
M.S. University of Central Florida, 2018

A dissertation submitted in partial fulfilment of the requirements
for the degree of Doctor of Philosophy
in the Department of Computer Science
in the College of Engineering and Computer Science
at the University of Central Florida
Orlando, Florida

Spring Term
2022

Major Professor: Pamela Wisniewski

ABSTRACT

Computational risk detection has the potential to protect especially vulnerable populations from online victimization. Conducting a comprehensive literature review on computational approaches for online sexual risk detection led to the identification that the majority of this work has focused on identifying sexual predators after-the-fact. Also, many studies rely on public datasets and third-party annotators to establish ground truth and train their algorithms, which do not accurately represent young social media users and their perspectives to prevent victimization. To address these gaps, this dissertation integrated human-centered approaches to both creating representative datasets and developing sexual risk detection machine learning models to ensure the broader societal impacts of this important work. In order to understand what and how adolescents talk about their online sexual interactions to inform study designs, a thematic content analysis of posts by adolescents on an online peer support mental health was conducted. Then, a user study and web-based platform, Instagram Data Donation (IGDD), was designed to create an ecologically valid dataset. Youth could donate and annotate their Instagram data for online risks. After participating in the study, an interview study was conducted to understand how youth felt annotating data for online risks. Based on private conversations annotated by participants, sexual risk detection classifiers were created. The results indicated Convolutional Neural Network (CNN) and Random Forest models outperformed in identifying sexual risks at the conversation-level. Our experiments showed that classifiers trained on entire conversations performed better than message-level classifiers. We also trained classifiers to detect the severity risk level of a given message with CNN outperforming other models. We found that contextual (e.g., age, gender, and relationship type) and psycho-linguistic features contributed the most to accurately detecting sexual conversations. Our analysis provides insights into the important factors that enhance automated detection of sexual risks within youths' private conversations.

I dedicate my dissertation work to my family and many friends. A special feeling of gratitude to my loving parents, Amir and Tahereh whose words always inspired and motivated me. My sister Afrooz, source of inspiration, who has always been there for me. My partner James who has always supported me with unconditional love during the challenges of graduate school and life.

ACKNOWLEDGMENTS

I would like to express my gratitude and appreciation for my advisor Dr. Pamela Wisniewski whose guidance, support and encouragement has been invaluable. Her insights and knowledge into the subject matter steered me through this research, and her support and encouragements made me the person I am today. And special thanks to my dissertation committee Dr. Munmun De Choudhury, Dr. Charles E. Hughes, and Dr. Ulas Bagci for their support, thoughtful comments, and recommendations on this dissertation. I would like to thank all the STIR Lab team at the University of Central Florida, who have supported me and this project. Furthermore I would like to thank the rest of the undergraduate research team for their collaborative effort during data collection. Finally, I would like to thank the participants who were generous with their time participating in this research and enabled this research to be possible.

This research is partially supported by the U.S. National Science Foundation under grants IIP-1827700 and 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.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION

Adolescents Digital Experiences and Sexual Risks

According to Pew Research [22], 95% of teens in the United States have access to a smartphone, and 45% say they are constantly online. As such, adolescent relationships are increasingly being mediated by their use of online technology. Over half (57%) of teens (ages 13-17) have started a new friendship online, and of the 35% of teens that have had a romantic relationship, 8% met their romantic partner online [167]. Another 55% of adolescents have used social media to flirt with someone they are interested in [167]. While the internet affords adolescents numerous opportunities to form new relationships and explore their sexual identities [270], it also has the potential to pose new sexual risks. For instance, the Crimes against Children Research Center estimates that 23% of youth in the U.S. have experienced unwanted exposure to pornography, and 9% reported receiving unwanted sexual solicitations online [142]. Over half of youth in the U.S. (ages 10 to 17) have received at least one online sexual solicitation (wanted or unwanted) in the past year[193]. As such, technology-facilitated sexual violence [123, 258], which manifests in non-consensual or unwanted sexting, sexual grooming, sex trafficking, and/or exploitation or abuse, has become a prevalent concern among internet users [123, 102]. For instance, the National Center for Missing and Exploited Children (NCMEC) [6] received more than 16.9 million child sexual exploitation reports in 2019, which included online child grooming, "sextortion" (i.e., threats to expose sexual images to coerce them to provide additional pictures, sex, or favors), and the engagement of children in sexual activity via the internet. Online grooming behaviors can also lead to sex trafficking, which is "the recruitment, harboring, transportation, provision, or obtaining of a person for the purpose of a commercial sex act" [7].

In 2018, internet-based sexual violence was thrust to the forefront of U.S. political legislation when

the Fight Online Sex Trafficking Act (FOSTA) and Stop Enabling Sex Traffickers Act (SESTA) bills [240] were signed and immediately put into effect. Together, these bills, for the first time in history, made online platforms accountable for sex trafficking facilitated via their platforms. Given the political landscape and the prevalence of internet-based sexual violence, the relevance and applicability of machine learning (ML) approaches for online sexual risk detection have become of critical importance when considering solutions for solving this dire societal problem. As such, researchers have begun trying to synthesize the state-of-the-art in computational ML for various online sexual risks [174, 138, 188, 266]. For instance, Tariq et al. [266] surveyed Computer Vision (CV) approaches for detecting skin dominance and nudity in digital images and videos for the purpose of combating adolescent sexting. They identified the need for more human-centeredness in designing and developing nudity detection algorithms to make them applicable for real-world deployment. In Chapter 2, we build upon this prior literature by conducting a human-centered systematic review of the computational approaches for online sexual risk detection within text-based and multi-modal data (i.e., text with images, videos, and/or meta data). We define computational risk detection as ML and other automated approaches that predict risk-related behavior [199].

A Case for Human-Centered Machine Learning

One may ask why we need human-centeredness in computational approaches for online sexual risk detection. Artificial Intelligence (AI) systems are used to make decisions across various human domains, and they impact people's lives in high-stake situations such as criminal justice [129], child welfare [246], and healthcare [65]. Past research has expressed concerns about these systems failing to account for limitations and/or uncertainties inherent in their predictions that may result in negative impacts to people's lives [280]. As such, "human-centered machine learning" (HCML) is an emerging sub-field of Computer Science that leverages computational AI expertise and human

knowledge from the social sciences to ensure that ML approaches to address societal needs and do no harm. Human-centeredness provides insight into the potential pitfalls and ethical considerations inherent in using technology to solve human problems that a purely computational lens lacks, providing a better understanding of how these models will perform in real scenarios and the potential impact these technologies will have on end users [139]. Moreover, Human-Centered Design (HCD) enables researchers to incorporate the perspectives of various stakeholders, helping them to construct a robust algorithm [245]; in our context, leveraging such a human-centered lens can facilitate reaching the goal of detecting, mitigating, and preventing online sexual risks.

With the rise of the #MeToo movement [125], sexual harassment and/or abuse prevention has also become a popular research topic within the SIGCHI and Computer-Supported Cooperative Work and Social Computing (CSCW) research communities, ranging from in-depth qualitative accounts of sexual victimization [20, 118] to computational approaches for sexual risk detection [262, 171]. Therefore, examining the online sexual experiences of modern-day adolescents is an important, and growing area of research within the adolescent developmental psychology literature [104, 143, 263] and the HCI research community [30, 43, 107, 185, 223, 299]. These online sexual experiences are a new manifestation of an age-old social computing phenomenon [83, 194, 247] that warrants our sustained attention. In Chapter 3, we build upon this prior research by analyzing social media trace data from an online peer support platform to understand how adolescents seek support and advice regarding their online sexual experiences to operationalize this knowledge for building models. In Chapter 4, based on the results of our systematic literature review which show that many studies rely on public datasets and third-party annotators to establish ground truth and train their algorithms, we build a human-centered data donation and annotation system to gather ecologically valid training datasets from youth. Then in chapter 5, we utilize this dataset to investigate machine learning approaches for the detection of online sexual risks. In Chapter 6, we conducted a meta research to investigate the implications of participating in a sensitive research for

both participants and annotators of social media data. Lastly in Chapter 7, we go over a summary of the overall findings, contributions and outcomes of this dissertation.

Research Questions

We are interested in answering the following over-arching research questions:

- **RQ1-Literature Review:** *What are the trends, gaps, and opportunities in the current literature of computational approaches for online sexual risk detection? How to address the gaps within the existing literature and provide recommendations for a research agenda that would advance beyond the state-of-the-art within this research domain?*
- **RQ2- Study 1:** *What are the key characteristics of online sexual experiences that adolescents seek support for?*
- **RQ3- Study 2:** *How could we gather ecologically valid and human-centered datasets for training machine learning models?*
- **RQ4- Study 3:** *Can we use human-centered machine learning to accurately detect these sexual risk experiences?*
- **RQ5- Study 4:** *Retrospectively, can we ensure the safety and well-being of our research participants and research team when studying online sexual risks by understanding the impact of flagging unsafe content on them?*

Dissertation Overview

In this dissertation study, we direct our attention toward a better understanding of the sexual content that adolescents exchange online and the development of algorithms that can detect these risks

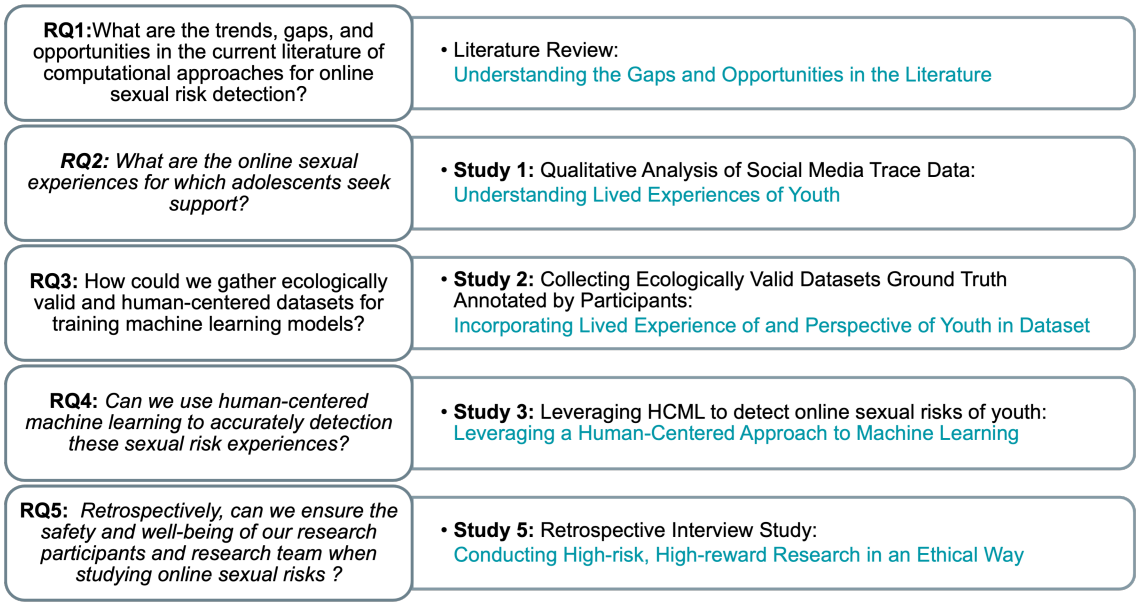

Figure 1.1: Dissertation Overview

accurately. An overview of this dissertation is displayed on Figure 1.1. In Chapter 2, we describe a human-centered systematic literature review that we conducted to answer RQ1 which summarizes the trends, gaps, and opportunities within the existing efforts that have been taken up in the area of computational sexual risk detection. Then, we introduce our framework for conducting our computational literature review and our methods in which we instantiated this framework to qualitatively analyze the literature. In Chapter 3, we address RQ2 by conducting a thematic content analysis of 4,180 posts by adolescents on an online peer support mental health forum to understand how adolescents seek support about their online sexual interactions, and what forms of support they seek. The goal is to develop a better understanding of how adolescents engage in online sexual experiences. In Chapter 4, we address RQ3 by presenting the design of a data donation platform for youth to gather ecologically valid datasets based on teens' real world social media data and contextualize their perspective on risk for labeling this dataset. The goal is to create contextual training datasets for adolescent online risk detection. In Chapter 5, we answer RQ4 by conduct-

ing experiments on the dataset gathered in Chapter 4 to build human-centered machine learning algorithms for various online sexual risk detection in adolescents private conversations. In Chapter 6, we answer RQ5 by presenting the results of the interviews with participants who donated their Instagram data and flagged their conversations for online risks and research assistants who annotated youth private conversations to investigate how flagging conversations for unsafe interactions impacted them. Lastly, in Chapter 7, we provide an overview of the outcomes from this dissertation.

CHAPTER 2: LITERATURE REVIEW

Citation: A. Razi, S. Kim, A. Alsoubai, G. Stringhini, T. Solorio, M. De Choudhury, P. Wisniewski, "A Human-Centered Systematic Review of the Computational Approaches for Online Sexual Risk Detection", at Proceedings of the ACM on Human-Computer Interaction (CSCW 2021).

In this chapter, we present our human-centered systematic literature review for computational sexual risk detection, which lead us to find gaps and opportunities to later on address some of these important issues in this dissertation.

Introduction

In the era of big data and artificial intelligence, online risk detection has become a popular research topic. From detecting online harassment to the sexual predation of youth, the state-of-the-art in computational risk detection has the potential to protect particularly vulnerable populations from online victimization. Yet, this is a high-risk, high-reward endeavor that requires a systematic and human-centered approach to synthesize disparate bodies of research across different application domains, so that we can identify best practices, potential gaps, and set a strategic research agenda for leveraging these approaches in a way that betters society. Therefore, we conducted a comprehensive literature review to analyze 73 peer-reviewed articles on computational approaches utilizing text or meta-data/multimedia for online sexual risk detection. We identified sexual grooming (75%), sex trafficking (12%), and sexual harassment and/or abuse (12%) as the three types of sexual risk detection present in the extant literature. Furthermore, we found that the majority (93%) of this work has focused on identifying sexual predators after-the-fact, rather than taking more nuanced approaches to identify potential victims and problematic patterns that could be used to

prevent victimization before it occurs. Many studies rely on public datasets (82%) and third-party annotators (33%) to establish ground truth and train their algorithms. Finally, the majority of this work (78%) mostly focus on algorithmic performance evaluation of their model and rarely (4%) evaluate these systems with real users. Thus, we urge computational risk detection researchers to integrate more human-centered approaches to both developing and evaluating sexual risk detection algorithms to ensure the broader societal impacts of this important work.

In the previous chapter, we introduced why human-centeredness in computational approaches for online sexual risk detection is needed. Traditional frameworks of computational ML risk detection focus on the 1) **Data** used to train the algorithms, 2) **Models** or algorithmic approaches for risk detection, and 3) **Evaluation** metrics for assessing how these models perform. Yet, we used a human-centered lens to conduct a systematic literature review of computational approaches for online sexual risk detection that led us to ask more nuanced research questions:

- **RQ1 (Data):** *Are the datasets ecologically valid for detecting the targeted risk for the desired user population?*
- **RQ2 (Models):** *Are the algorithmic models grounded in human theory, understanding, and knowledge?*
- **RQ3 (Evaluation):** *How well do the algorithms perform, both computationally and in meeting end users' needs?*
- **RQ4 (Application):** *What system artifacts were developed, and what were the outcomes when deployed in real-world settings?*

To answer these research questions, we conducted a systematic literature review which analyzed 73 peer-reviewed papers published between 2007 and 2020. We performed a comprehensive liter-

ature search to identify any computational approaches for online sexual risk detection within text-based and multi-modal data. In our literature review, we broadly considered all types of online sexual risks that may result in mental or physical harm, including sexual violence/abuse, sexual harassment, sexual grooming, and sex trafficking. We qualitatively coded these articles using a human-centered lens that assessed the ecological validity of the data being used to train the algorithms, the algorithmic approaches being used, the metrics for which to assess the quality of these models, and whether and how these models were deployed in real-world settings.

Overall, we found that most papers proposed algorithms for detecting sexual predators (75%) after the sexual violence occurred (93%) using public datasets (82%). These findings imply that there is a need for approaches that help prevent victimization and to detect other types of sexual risks, such as sexting. We identified that most (52%) approaches relied on datasets that are not representative of end users, and are annotated by third parties without adequate background on the subject (31%). The takeaway from this finding is that it is crucial to have realistic datasets with high quality annotations for training models that take into account the perspectives of the individuals who experienced the risk as well as experts and clinicians to do not exclusively rely on researchers and volunteers. We found that most (60%) models do a binary classification of users assessing whether they are a predator or not; there is a need for approaches that take into account patterns and changes at the conversation level so that detecting the risky content will be more effective and meaningful. The results of our review for evaluation methods illustrate that the main (81%) focus is on computational evaluation of the method and that there are only a few (8%) user studies to evaluate the developed technology. In addition, most (93%) of the studies proposed algorithms but did not integrate these algorithms in a real system that could be used by stakeholders to identify and mitigate sexual risks. Our research makes the following novel contributions to the HCI, HCML, and ML research communities:

- A conceptual framework for systematically reviewing computation risk detection literature using a human-centered lens (Figure 1)
- An in-depth synthesis of the current state-of-the-art and trends in computational approaches for online sexual risk detection
- Identification of the potential gaps within the existing literature and recommendations for a research agenda that would advance beyond the state-of-the-art within this research domain

In the following section, we describe the existing efforts that have been taken in the area of computational sexual risk detection. Then, we introduce our human-centered framework for conducting our computational literature review and our methods in which we instantiated this framework to qualitatively analyze the literature.

Background

Prior reviews of computational approaches for sexual risk detection were typically couched more broadly in the examination and detection of online abuse and cyber-agression. For instance, a review by Mishra et al. [191] presented the computational approaches for detecting online abuse, including racism, sexism, personal attacks, toxicity, and harassment. They took a coarse-grained definition of abuse as“any expression that is meant to denigrate or offend a particular person or group", in their review of the literature. Similarly, Mladenovic et al. [198] reviewed papers focused on detecting cyber-aggression, cyberbullying, and cyber-grooming. The review by Mladenovic et al. [198] took a generalized view in reviewing papers that included any risks (including sexual) within aggressor–victim relationships in online platforms. In both reviews [191, 198], they found that deep learning models, such as Convolutional Neural Networks (CNNs), achieved better accuracy than traditional machine learning models.

Mladenovic et al. [198] also found that the most useful features are word and character embeddings. They also mention that English is the pre-eminent language studied from the perspective of all the three reviewed risks and there is a gap of having datasets in different languages. These researchers made notable contributions by synthesizing the literature and identifying potential research gaps using a primarily computational lens.

From a human-centered perspective, we argue that sexual risks are a unique form of violence and that reviewing literature for different risk phenomenons (e.g., cyberbullying vs. cyber-grooming), under the assumption they are similar, may lead to false conclusions or unintended and/or negative consequences for sexual violence victims. Different risk types have different characteristics of victimization that should be taken into consideration; for example, cyberbullying and cyber-grooming are distinct risk types, even though they share a generalized definition as an attack directed to harm a victim [198]. Cyberbullying includes an intentional and aggressive act against someone or a group of people while cyber-grooming happens between a sexual predator and victim to gain the victim's trust for the purpose of sexual abuse [198]. Each of these risks deserves careful attention and exploration. Therefore, we reviewed papers specific to sexual risk detection, which includes sexual grooming, sex trafficking, and sexual harassment and/or abuse.

A couple of reviews have been conducted on online sexual risk detection in the specific context of sexual grooming and identification of child sexual predators. These reviews coincided with the 2012 Sexual Predator Identification competition ran by PAN[1] using the PAN-12 dataset. This competition aimed to provide researchers with an initial benchmark for comparing different methods of detecting cyberpeodophiles or sexual groomers by using PAN-12, a large dataset of chat logs between convicted sex offenders and volunteers posing as children (created by Perverted Justice [134]). Inches and Crestani [134] published a review of approaches that were taken during the

[1] A benchmarking activity on uncovering plagiarism, authorship and social software misuse http://pan.webis.de

competition. In their review, they discussed the submissions' pre-filtering techniques, features, the classification models, and evaluated the computational approaches submitted by 16 teams. The top five teams who were able to identify sexual predators attained a higher accuracy using Support Vector Machine (SVM) algorithms. This research provided a framework for benchmarking the computational approaches for sexual predator identification from a technical standpoint. Ngejane et al.[206] continued the effort by reviewing 10 additional papers that used the PAN-12 dataset that were published after Inches' and Crestani's [134] review and found that most of the reviewed approaches used supervised models; among these supervised models, they confirmed that SVM yielded the highest accuracy (98%) in detecting child sexual predators in the PAN-12 data.

A common theme among these prior syntheses of the computational literature for both generalized and/or sexual risk detection is that they focused predominantly on the algorithms themselves in terms of the standard ML metrics for benchmarking performance. Although quantitative measures are necessary for evaluating the algorithms' performance, quantitative measures are not sufficient in determining if the models perform well from the point of view of stakeholders and end users. Hence, these computationally-focused reviews suffer a disconnect between the impact on the functionality of the algorithms and the social interpretations needed to assess quality in the broader sense of the problem context [35]. As Baumer [35] suggests, these disconnects can be addressed by incorporating human-centeredness to evaluate the approaches with deeper interpretations behind algorithmic inferences that impact real-world use [35]. Unlike the previous works, our review leverages this human-centered lens to synthesize the computational literature on detecting online sexual risks. Sexual victimization is a social issue that requires the integration of personal, social, and cultural aspects for designing and developing intelligent systems that understand this socially complicated issue. Thus, our work aims to bridge the gap between social needs and computational perspectives by using a human-centered lens to synthesize the computational risk detection literature in this domain. In the next section, we introduce our human-centered lens for reviewing

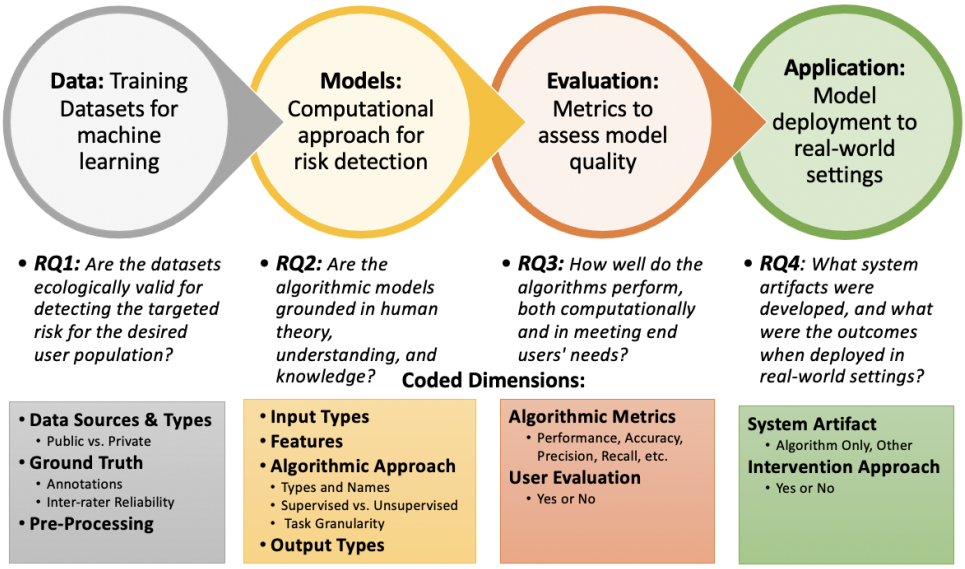

Figure 2.1: Human-Centered Lens for Computational Risk Detection Systematic Literature Reviews

computational risk detection in the context of sexual risk detection.

Applying a Human-Centered Lens to Computational Risk Detection

We present four main components of computational ML systems (i.e., Data, Models, Evaluation, and Application) and demonstrate how a human-centered lens can be applied for analyzing computational risk detection research (Figure 2.1). Traditionally, ML researchers focus on data, models, and computational evaluations of these models. These are the main components of ML which every systematic literature review should take into account; yet, these components should also consider the human-context in which these models intend to be deployed. To do this, we synthesized relevant bodies of work across the fields of Human-Computer Interaction (HCI), HCML, and ML to create a conceptual framework in which to conduct our literature review. This framework is one of the contributions of this paper and also served as a theoretical lens for grounding the

analyses of the reviewed papers within the domain of online sexual risk detection. While we apply this framework to the specific context of online sexual risk detection, it can also be generalized to other forms of computation risk detection that could benefit from a human-centered perspective.

The ubiquitous use of systems to produce risk predictions have consequences that impact people's lives [209]; ML/AI systems have been used in decision making systems in various contexts [14], such as child welfare [246], that affect people's lives in profound ways. Recently, researchers across multidisciplinary fields, including HCI, ML, public policy, and humanities have come together to address the gaps between computational AI systems and societal needs. This had led to the paradigm shift toward HCML, which tries to address a system's inability to improvise according to context and human characteristics such as perceptions, emotions, intentions, and social contexts [38, 296] to keep humans at the center of the design process by taking into account stakeholders needs. Yet, researchers have varied on a formal definition for human-centered computing; Kling and Star expressed that "there is no simple recipe for the design or use of human-centered computing" [151]. Chancellor et al. [63] argued that HCML is a growing interdisciplinary field that there are no formal definitions for, so they sought to understand the definition of "human" in regard to HCML for mental-health. Similarly, we apply a human-centered lens broadly for reviewing computational approaches for online sexual risk detection.

Importantly, there are key differences that distinguish computational risk detection from general computational ML predictions that need to be considered when reviewing the literature. Risk is a subjective concept, which makes it difficult to operationalize [222]. Thus, it is important to look at the specific problem and the context around which risks occur. As the goal of detecting risk is typically to protect people from harm, inaccurate risk detection could be potentially life-altering, creating real-world ramifications for both potential victims and alleged predators. When applying ML to risk detection, false positives or/and negatives may have adverse consequences for users. As such, computational risk detection is a high risk, high reward research problem, which requires

extra scrutiny. In Figure 2.1, we reiterate our high-level research questions and map these questions to dimensions of the literature that coded for in our systematic review of the sexual risk detection literature. In the sections below, we describe each component of our human-centered framework in more detail.

Considerations for Data Using a Human-Centered Lens

Data is defined as sets of instances for building or evaluating ML algorithms, and label is a value or category assigned to each data instance and served as the target for the algorithm [199]. Given the fact that data is the foundation of algorithmic development, it is important that the data matches the real world users' context [119]. Knowing whether the data is the true representation of user behavior is important, so computational risk detection research should consider the motivations behind how the dataset was created, as well as the characteristics of the data itself. The nature of the risks that happen on various platforms is distinct, as each online platform has its own characteristics and affordances [178]. For example, the nature of the risks in public posts on Twitter may be different than the nature of private conversations on Facebook. Researchers also need to review the mechanisms or procedures of the data collection and ground truth annotations to examine if there were clear explanations behind the process as well as steps to validate such procedures. Researchers should also analyze the types of data included in the dataset. When it comes to detecting risks, important indicators, such as the context of the relationship [231] or time of communication [172] may be indicative of risk. Since risks are context-driven [222], having multi-modal data points could help improve model performance.

Other data considerations may include privacy policies, consent processes (if data is collected from human subjects), ethical considerations, and potential sampling bias. For instance, the frequency and the nature of risks that occur in private spaces are different than the risks that happen

in public spaces [303]. Having transparency about the data collection process would help users and researchers to identify any types of biases or data privacy issues, to ensure correct use and distribution of the data. For instance, if researchers are using public datasets to detect risks, then the nature of risks will be limited to these public contexts. Additionally, how the data is labeled for the ground truth also matters, as labels for training directly affect the models' learning of a particular phenomenon and the results of ML algorithms [242]. For instance, Kim et al. [148] found that third-party annotators had a significantly different perspective on bullying-related risks compared to the victims themselves. Annotators' background and expertise influence how the data is annotated for ground truth; thus, annotators should be selected carefully and their background should be stated clearly [11]. Essentially, such considerations interrogate whether the data being leveraged is ecologically valid for detecting the targeted risk for the desired user population. When there are multiple people labeling the same dataset, it is important to measure the Inter-rater Reliability (IRR) with at least a subset of the corpus as well as formulate ways to resolve any annotation disagreements between the annotators [169].

Considerations for Computational Models Using a Human-Centered Lens

Models, from a computational perspective [199], are the artifacts that encode decision or prediction logic trained from the training data, the learning program, and frameworks. The model learns from a set of features; a feature is a measurable property or characteristic of a phenomenon being observed to describe the instances. From a human-centered perspective, it is important to see if the computational models being built are human-informed and evidence-based in terms of drawing from risk models based on human theories [35]. Baumer et al. [35] introduced human-centered approaches to machine learning that leverage theoretical frameworks derived from and validated within the social sciences, such as psychology and communications. So, the main question to be asked for reviewing computational risk detection research revolves around "Are the algorithms

grounded in human theory, understanding, and knowledge?". When reviewing computational risk detection models, it is important to look into explanations and the transparency of the model to understand if it could be accountable for making fair and unbiased decisions by utilizing human-centered approaches [187]. Reviewing explanations for different parts of the models includes, but is not limited to, identifying the types of algorithms or techniques that were used and why, how the algorithms were trained, what the input characteristics were, training parameters, fairness constraints and potential biases, features, and if the intended use of the output of the model and the choice of output structure was stated [126]. More importantly, depending on the task that the model is trying to do, especially in the area of risk detection, the outcomes of the model could be reviewed from different human angles; for instance, looking at the timing of the detection to understand if the model is capable of detecting a risk before or while it happens, or if it needs full data to detect the risk after-the-fact [172].

Human-Centered Considerations for Model Evaluation

Once a machine learning model is developed, one must test the model to evaluate the performance and speculate the usage of the model. From a computational perspective [199], generalization error is defined as the expected difference ratio between the real conditions and the predicted conditions of any valid data. Usually, computational scientists test the models based on computational metrics such as accuracy, precision, or recall. Reporting precision and recall is important based on the application that the risk detection system will be used. Although such metrics provide a picture of the capabilities of the model, they are far from sufficient for evaluating the model. From a human-centered perspective, researchers could ask "How well do the algorithms perform computationally and in meeting end users' needs??" HCI design methods such as user studies could be used to evaluate the model from the users' perspective. User studies could provide valuable insights from stakeholders of the system including end-users. Yet, testing is important in the life

cycle of deploying a machine learning system. Zhang et al. [307] summarized techniques for testing ML systems from a computational perspective such as testing properties (e.g., correctness, robustness, and fairness), testing components (e.g., the data, learning program, and framework), testing workflow (e.g., test generation and test evaluation). They mentioned before deploying the model online, conducting offline testing, such as cross-validation, to make sure that the model meets the required conditions is necessary. After deployment, predictions of new data can be analyzed via online testing to evaluate how the model interacts with user behaviors.

Methods

Below, we describe how we scoped our literature search and systematically reviewed the articles included in our dataset.

Systematic Literature Search

For our initial literature search, we identified five electronic databases that included computational and interdisciplinary research on sexual risk detection (i.e., IEEE Xplore Digital Library, ACM Digital Library, ScienceDirect, Springer-link, and ACL Anthology) to ensure comprehensive coverage of the relevant literature. We used combinations of the following keywords: sexual predator, sexual predation, sexual risk detection, sexual abuse detection, sexual grooming, sexual assault, online grooming, cyberpedophile, pedophile, paedophile, sex trafficking, predatory conversations. We included words like detection, recognition, and machine learning to find computational articles, as opposed to articles that studied the phenomenon itself from a more qualitative perspective (e.g., focus groups or interview studies). We explain our scoping criteria and relevancy coding next.

Scoping Criteria and Dataset Creation

Our initial search resulted in 296 unique papers. Next, we examined the paper title, abstract, keywords, results, and conclusions to identify relevant studies that met the following inclusion criteria

- The study was a peer-reviewed published work.
- The study was published between 2007 and 2020. We included papers that were published in these years, but there were not any papers that met our inclusion criteria before 2007.
- The study focused on sexual risk detection (our definition of sexual risks includes sexual predation, sexual grooming, sexual assault, sexual abuse, sex trafficking, and sexually abusive conversations).
- The study contained a computational/algorithmic approach or a system architecture on text and multi-modal data (including Natural Language Processing, Machine Learning, etc.)

We marked a paper as relevant if it met our relevancy criteria, which resulted in 57 relevant articles. Then, we cross-referenced the citations of these relevant papers to identify additional papers to include that may not have been included in our initial database search. This cross-reference exercise resulted in 116 unique papers (62 after removing duplicates from our initial search) in which 15 papers met our inclusion criteria. We did one more iteration of this search process, which identified two additional relevant papers. Having reached an apparent saturation point, we concluded our search with a final set of 73 articles for our review.

The primary reason papers were excluded (around 80%) was because the article did not specifically address sexual risks. These papers included cyberbullying detection [241], authorship attribution [138], and general privacy [32] and forensics papers [34]. The second most common reason

for exclusion (around 15%) was that the study did not include a computational approach for sexual risk detection, such as papers that qualitatively assessed users' online sexual risk experiences [231]. We also excluded papers on image and video sexual risk detection papers because Tariq et al.'s [266] recent literature review on these approaches. However, we did include multi-modal approaches that included textual data. Next, we describe how we synthesized the literature.

Data Analysis Approach

We leveraged the human-centered lens proposed above (Figure 2.1) to code the 73 relevant papers. Our final codebook is presented in Table 2.1. We used an iterative process to refine our codes and allowed codes to overlap and be double-coded. Three coders labeled the same 15% of articles, and we calculated Fleiss's Kappa IRR [97], which is an extension of Cohen's kappa for three raters or more. This resulted in Fleiss's Kappa ranging from a substantial (0.70) to a complete agreement (1.00) [97] for all codes across dimensions. To resolve conflicts, the researchers discussed the articles until a consensus was reached and updated the operationalization of the codes for consistency. Then, the remaining articles were divided among the three coders. The first author reviewed the final codes to identify emerging themes, patterns, and potential gaps within the literature.

The definitions of our coded dimensions, local research questions, and grounded codes that emerged from our data are shown in Table 2.1. For instance, for Ground Truth, we coded for the annotation of the datasets and the humans' involved in the annotation process. Our codes include "Outsiders" (Someone other than the victim or the one that shared the story), "Insiders" (Victims or who shared the story), "Auto" (Used an automatic approach to label the data), or "Existing" (The paper used an existing labeled dataset). Task Granularity referred to the detection task granularity level for the studies that are on conversations, since there are several ways that researchers can implement risk detection on conversations (i.e., if the paper detects the "Level" of risk in a conversation,

Table 2.1: Codebook

Categories	Dimensions	Codes (Subcodes)
Overall Characteristics	**Risk Type**: *What are the risk type and the key aspects of the risks that researchers are detecting?*	Sexual Grooming (75%), Sex Trafficking (12%), Sexual Abuse/Harassment (12%)
	Timing of Detection: *When is the timing of the risk detection?*	After (93%), During (38%), Before (5%)
	Target Person: *Who is the person that the system tries to detect? What is the age range of the target person?*	Perpetrator (80%) (Child, Adult, Posing as Child), Victim (17%) (Child, Adult, Posing as Child)
Data	**Data Source**: *What is the data source and data type? What is the privacy level of the dataset?*	**Dataset Source**: Perverted Justice (30%), PAN-2012 (22%), Combined Chat datasets (14%), Social Media (11%), Advertisements (11%), Queries (5%), Private Chat data (4%), SafeCity (4%), Games (3%), Forums (3%), Anonymous Platforms (3%), Blogs (3%), Generated by Participants (1%) **Data Types**: Text (64%), Meta Data (23%), Images/Video (11%) **Privacy Level**: Public (82%), Private (15%)
	Ground Truth: *How was the data annotated for training datasets? Did they report annotators' IRR for more than two annotators?*	**Annotators**: Existing Labels (44%), Outsiders (33%) (Researchers (26%), Moderators (3%), Clinical (1%), Crowdsource (1%)), Automatic Approach (23%), Insiders (5%) **Annotators IRR**: No (95%), Yes (5%)
Models	**Feature**: *What were the features used in the development of the model?*	Textual (85%), User (27%), Time/Location (19%), Semantic (16%), Style (15%), Behavioral (14%), Keyword Extraction (14%), Syntactic (11%), Sentiment (10%), Images (7%), Network (5%), Relationships (4%), Topic Modeling (4%)
	Algorithmic Approach: *What machine learning model(s) were used for the task? What is the detection task granularity for the studies that are on conversations?*	**Approach Types**: Machine Learning (83%) (Traditional (66%), Deep Learning (15%), Ensemble(2%)), Handcrafted/Rule-based (18%), System Architecture (7%), Graph/Network-based (7%) **ML Model Type**: Supervised (72%), Unsupervised (13%), Semi-Supervised (5%) **Algorithm Name**: Support Vector Machine (37%), Bayes (22%), Neural Networks (21%), Regressions (14%), etc. **Task Granularity**: Users (36%), Conversation Type (21%), Patterns (12%), Lines/Parts of Conversations (11%), Risk Levels (3%)
	Output Type: *What is the format of the model output?*	Binary (60%), Multi-class (33%), Clusters (10%), Predatory Stages (7%)
Evaluation and Application	**Evaluation**: *How was the model evaluated? Did they do user studies to evaluate the model by stakeholders?*	**Algorithmic Metrics**: Accuracy (53%), F1 (37%), Precision (23%), Recall (19%), etc. **User Study**: No (96%), Yes (4%)
	Artifact and Intervention: *What is the final artifact developed? Did the artifact provide any interventions of risk mitigation in addition to detecting risk?*	**Artifact**: Algorithm Only (86%), Chat System (1%), Forensic Investigation Tool (1%), Mobile App (1%) **Intervention**: No (93%), Yes (7%)

identifies the "Lines" of a conversation that are risky, identifies risky or predatory "Patterns" in conversations, or identifies Users as predators or victims in a conversation, or classifies the whole

"Conversation" into different categories). Next, we present our results regarding our analysis of the categories according to the coodebook and findings from the literature.

Results

In this section, we present our findings based on the 73 papers that we reviewed. We organize and present the results by our code dimensions in Table 2.1.

Overall Characteristics of Articles

Below, we report the descriptive characteristics of the articles in our dataset, including the risk types studied in the literature over time, timing of detection, and target (i.e., person) of risk detection.

Table 2.2: Risk Types

Risk Types: Definition	**Count (Percent)**	**References**
Sexual Grooming: Detecting sexual grooming of an online predator. Grooming is a process to approach, persuade, and engage a child, the victim, in sexual activity by using Internet as a medium [208].	55 (75%)	[146, 92, 177, 219, 174, 192, 250, 49, 305, 46, 51, 237, 175, 15, 309, 244, 114, 214, 213, 86, 135, 86, 164, 224, 225, 47, 99, 27, 271, 61, 279**?** , 67**?** , 138, 163, 91, 109, 74, 272, 217, 68, 45, 136, 189, 24, 184, 188, 155, 218, 238]
Sex Trafficking: Is the process of recruitment, harboring, transportation, provision, or obtaining of a person for the purpose of a commercial sex act [7].	9 (12%)	[288, 254, 132, 131, 133, 130, 168, 158, 251]
Sexual Harassment/Abuse: Includes online shared content of a range of sexually aggressive or harassing content, sexual assault, gender violence, sexual violence such as stories shared in the hashtag MeToo movement [123].	9 (12%)	[144, 108, 302, 171, 120, 260, 243, 145, 262]

Sexual Risk Detection Types Overtime

As illustrated in the Table 2.2, the research papers that we reviewed mainly considered four kinds of sexual risks: sexual grooming (N=55, 75%), sex trafficking (N=9, 12%), and sexual harassment and/or abuse (N=9, 12%). As shown in Figure 2.2, we observed some notable trends. Starting from 2007, with the rise of using social media and online chat rooms [210], researchers start working on combating the emerging problem of sexual grooming. In 2012, we saw a spike in the literature, likely due to the PAN12 competition for predator detection [134]. Concurrently, sex trafficking detection literature emerged in 2012, which coincided with the announcement of the Obama administration efforts to combat human trafficking [1]. Meanwhile, we saw the rise of the #MeToo movement at the end of 2017 with social media users sharing sexual harassment self-disclosures that went viral [125]. Consequently, this coincides with the computational research on sexual harassment and/or abuse that emerged in 2018. Further, the significant increase in sexual risk detection publications in 2019 highlights the critical importance of computational sexual risk detection and the need for systematic review that synthesizes this body of research in a way that can create a cohesive research agenda moving forward. There seems to be a slight decrease of literature in 2020, this might have been caused by several underlying factors. We will further discuss these factors in our discussion section.

This trend in the past literature represents the algorithmic advances on sexual grooming detection which is an important issue to tackle. But also demonstrates the overlooking of critical sexual risk types that should carry the same weight, if not more, in some cases, such as unwanted sexting and sexual abuse. The long-lasting impacts on victims, as well as the distinctive characteristics of relationships that engender sexual risk (often someone close to the victim like a family member) [231] together underscore a necessity to close this gap in examining sexual abuse online.

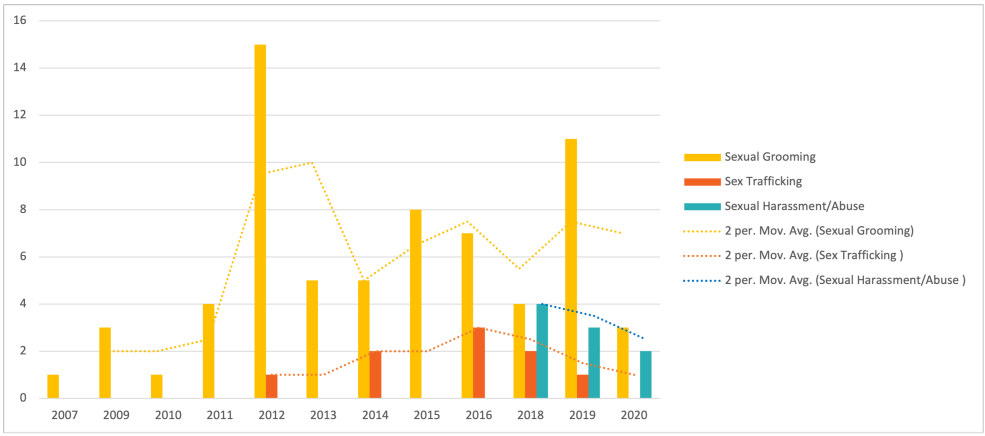

Figure 2.2: Number of Publications by Risk Type Over Time

When Risk Detection Occurs

Most algorithms in our review focused on after-the-fact risk detection (N=68, 93%), which examines ways to detect the risk with an underlying precondition after online risky behavior has occurred. For instance, Ringenberg et al.'s [237] machine learning model to differentiate contact-driven and fantasy-driven sexual solicitors were based on the dataset of complete conversations. Yet, there was a smaller subset of studies (N=28, 38%) that were capable of detecting risks during the occurrence through detecting predatory patterns and risk levels within a conversation [136, 164**?** , 131, 168, 155, 184, 224, 243, 250, 61, 114, 67, 214, 84, 272, 47, 163, 45, 74, 91, 49, 305, 192, 51, 109, 92, 238]. For instance, Cano et al. [61] sought to detect three online grooming stages: 1) Trust Development; 2) Grooming; 3) Seek for physical approach, while sexual grooming is happening. Only 4 (5%) papers detected sexual risks before and during the event [177, 190, 219, 158]. MacFarlane et al. [177] detected personal information in children's online chats and proactively blocked private information from being sent to prevent victimization. Kostakos et al. [158] built a model to predict risk factors of users that depict the likelihood of being drawn to online sex work and illustrated a potential methodology that could be used to identify

ones with high-risk factors. The way that the data is given to a computational model as an input affects the fact on when the prediction result of the model. In summary, the automated models for sexual risk detection of the past literature have generally been based on after the risk occurs. While detecting online risky interactions after-the-fact could help law enforcement agencies detect sexual predators, at this point it may be too late to prevent victimization.

Whom is the Object of Sexual Risk Detection

The majority of the papers focused on identifying predators (N=59, 80%) or conversations that included predators. Another 13 (17%) papers [177, 254, 131, 144**?** , 260, 158, 262, 171, 302, 108, 145, 120] focused on identifying victims. The focus of these papers was mostly to identify social media self-disclosures about sexual harassment and abuse (N=11, 15%). For example, Hassan et al. [120] leveraged the hashtag #MeToo movement, where women shared their stories of sexual violence in social networks, to collect and to classify the sexual harassment reports among the posts.

Age Range for Sexual Risks Detection

Most computational approaches to sexual risk detection were based on predators who were adults and victims who were adult volunteers posing as children (predator/pseudo-victim) (N=40, 55%) given the prevalent use of the Perveted Justice and PAN-12 datasets. One paper [**?**] created a chatbot posing as teens to talk to pedophiles. There were 20 (27%) papers that either did not specify the age of predators or victims or used data from adults. As an instance of not specifying age, Karlekar et al. [144] focused on identifying sexual harassment posts on SafeCity dataset without differentiating users by age. Although some posts were about the child abuse disclosures, they do not directly identify child victims. In fact, in some cases, an adult shared a story about a child.

These papers were indifferent to main dissimilarities between data from under-age users and adult users. Only a few (N=9, 12%) papers [177, 136, 254, 288, 164, 131, 260, 214, 190] focused on underage victims, which were mostly proposing chat systems or crime investigation tools, but not proposing models on chat conversations. A unique study in terms of doing an effort to focus on youth's data is the study by Roy et al. [243] which used hypothetical situations and created a dataset by recruiting youth and gave them scenarios to create abusive text messages given each scenario.

Assessing the Ecological Validity of the Data (RQ1)

In this section, we present our findings regarding the data and ground truth annotated training datasets used for sexual risk detection.

Data Type: Primarily Focused on Text

The majority of the papers (N=47, 64%) relied on text data for their risk detection models, (media types for each reference is displayed in Table A.1 in the Appendix). Although the datasets for these papers include metadata, for example, PJ has timestamps of the messages, these papers did not utilize this additional data. There were some papers (N=16, 21%) that incorporated metadata in addition to text such as time or user profile data. For example, Potha et al. [224] used time series modeling to reveal conserved temporal patterns or variations in the strategies of predators in the PJ dataset. As such, approaches that focused on text often did so in the absence of images (N=66, 90%) or other types of multi-modal data. A few papers (N=7, 10%) considered images, text, and meta-data. Most of these papers (86%) were trying to detect sex trafficking from online advertisements, and there were no papers using images and multi-modal data for sexual risk detection within conversations. We found 2 (3%) papers that solely used metadata, such as profile features. Only one paper [272] utilized both text and images to detect adult content on Facebook

posts. Overall, as we included any articles that included text plus other data types, we identified a gap in the literature for analyzing multi-modal and metadata such as user and temporal features in order to acquire the context of the online interactions to build effective detection models for sexual risks.

Data Sources: Mostly Publicly Available and Public-facing Datasets

We analyzed papers in terms of their datasets' source, type, and privacy. Our analysis uncovered that papers mostly worked on public datasets (N=60, 82%), and only a few (N=11, 15%) were using private datasets [136, 219, 271, 243, 260, 262, 68, 15, 74, 189, 145]. Table A.2 in the appendix illustrates the datasets of the papers examined in this literature review. Private datasets are mostly scraped from Social Media (e.g. Twitter), but some kept the platform name anonymous and some were private chat data (e.g. Whisper [271]). Some papers used several datasets for their analysis, for instance, Pandey et al. [214] used social media data, blog posts data, and data from forums and public chats.

Next, we discuss the dataset type, name, and source based on prevalence of datasets. We identified that based on risk types that these papers are trying to detect, the sources of the datasets have commonalities. The most popular public datasets used in the literature for identifying sexual groomers include Perverted Justice (PJ) dataset [100] (N=22, 30%) and dataset from PAN-2012 competition (N=16, 22%) [134]. A few papers (N=10, 14%) used combined chat datasets with PJ chat data; for instance, Rangel et al. [228] used the PJ dataset and the dataset for author profiling at PAN 2013. Pranoto et al. [225] used scripts from www.literotika.com which contains sexual conversations shared by adults in a legal manner. Overall PJ and PAN-2012 were the only public datasets that helped researchers advance computational solutions for identifying sexual predators.

Moreover, some studies (N=8, 11%) used data from social media, for instance, Suvarna et al. [262]

sought to identify sexual assault victim-blaming language on Twitter scraped posts. They justified their decision of creating a dataset using Twitter by explained that other platforms such as Reddit's or Facebook's community affordances may not provide a structure for people to voice their actual opinions for sensitive topics such as blaming the victim due to the presence of moderators. Datasets from online advertisements were all used by sex trafficking papers (N=8, 11%) except one paper on sex trafficking [158]. This paper, utilized data from a popular European adult dating forum to collect data for each user's profile page and qualitative data on users' self-reported behavior regarding paid sex. In addition to the dominant trend of a few public datasets and social media data, there were noteworthy papers that used different datasets; queries in P2P systems or networks (N=4, 5%), private chat datasets (N = 3, 4%), and a public dataset named SafeCity (N=3, 4%) were seen in papers. All sexual harassment and abuse detection studies are mostly based on data scraped from social media and the SafeCity dataset, with the exception of a study by Roy et al. [243]. For instance, Khatua [145] extracted 0.7 million tweets with the hashtag "MeToo" for the social movement against sexual harassment. Due to the lack of publicly available datasets that are representative of users and the difficulties to collect real digital trace data because of the sensitive nature of this problem space, Roy et al. [243] recruited participants to generate abusive text messages that could be used for classification. The participants were given abuse scenarios and asked to create corresponding text messages. One of the limitations related to this dataset was that there is not enough data to capture all the features related to dating abuse. Overall, we observed some good practices among the literature for justifying their choice of dataset, but the datasets were heavily skewed towards the PJ dataset that is not based on real-world users' conversations. Therefore, these conversations cannot be considered ecologically valid for the basis of risk detection. In addition, most of the time risks occur in private conversations [303], the primary use of publicly scraped datasets limits their usefulness in addressing the actual problem.

Ground Truth Annotations: Reliance on Third-party Annotators

Our finding on how papers provided ground truth labeling for ML training datasets and who annotated the data unpacks that most of the studies relied on existing annotations that were done mostly by researchers without relevant domain expertise, this is illustrated in Table A.3 in the appendix. We found that a noteworthy proportion of the papers that we analyzed were based on existing labeled datasets (N=32, 44%). These datasets were mostly either PJ or the PAN12 dataset (N=23, 32%). Of the research papers that we reviewed, there are 45 papers that labeled their own dataset, most of which relied heavily on third-party (N=24, 33%) rather than the person who experienced the sexual risk. We further examined the papers that relied on third-party annotators for the data, particularly focusing on whether the papers explicitly included descriptive data about the annotators such as their backgrounds, expertise, and if any types of training was provided to them. Most third-party annotators were researchers (N=19, 26%) that did not appear to have special expertise on the matter, and only one paper had annotators that were clinically informed or experts. Chowdhury et al. [108] expressed that they had three independent annotators from Clinical Psychologists and Academia of Gender Studies for annotating their entire dataset of self-disclosure of sexual harassment tweets. The authors also provided their annotation guidelines about the types of posts that were considered sexual harassment and the annotation process.

There were only a few papers (N=4, 5%) that took into account the perspective of the individuals who experienced the sexual risks. These papers share a common characteristic of being on social media posts or tweets of self-disclosures of sexual harassment [260, 108, 145]. The other paper by Kostakos et al. [158] selected a European adult dating forum for data collection and used social data mining to collect quantitative data for each user's profile page and covert online ethnography to collect qualitative data (interviews) on users' self-reported behavior regarding paid sex. They labeled user profiles based on the interviews conducted with users about their tendencies to buy

and/or sell sexual services in the forum. This dataset was used to predict the risk factors of a larger poll of users. We found that only 4 (5%) papers [243, 171, 108, 155] reported Inter-rater Reliability (IRR) when they had more than one annotator to measure the quality of agreement between annotators. Liu et al. [171] and Chowdhury [108] articulated the annotation procedure and used Cohen's kappa coefficient of inter-rater agreement for labeling sexual harassment stories, while Roy et al. [243] stated using Light's Kappa; an inter-rater agreement statistical consistency measure. Kontostathis et al. [156] reported their IRR using Holsti's method for annotating Perverted Justice's conversations.

In addition, some papers used automated methods such as keyword match or classification software for the annotation of their datasets (N=17, 23%). For instance, Michalopoulos et al. [189] used a software for document classification that performs instant classification of a given text. Kontostathis, et al. [155] built a keyword-based software named ChatCoder to label and analyze predation chats and to identify a luring category for each. To make it more suitable to the online context, they adopted a simplified version of the Luring Communication Theory (LCT), a framework proposed by Olson et al. [208] that describes the process child sexual predators use to lure their victims into a sexual relationship. The study provided extensive details on the labeling process such as how the codebook for the labels was developed as well as the coding process stages. In summary, we found the reliance on existing data labels among the literature in this area and the ones who labeled the data were mostly by researchers.

Models Grounded in Human-Centered Theories and Knowledge (RQ2)

In this section, we provide our results of the review about models and their characteristics including features used by models, approach type and model selection, and model output.

Feature Selection: Textual versus Behavioral Features

The most commonly used features by papers were textual or lexical features (N=62, 85%), which could be drawn from raw text or statistics of the input text, however, theory or behavioral driven features are rare among the literature (illustrated in Table A.4 in appendix). The textual features include n-gram use, character modeling, bag-of-words (BoW) models, term-frequency-inverse document frequency (TF-IDF), word embeddings, skip grams, Hashing Vectorizer, the length of the text, count/ratio of "emoticons", count/ratio of profanity, and number of pronouns. For instance, some approaches used the number of reframing verbs (e.g., teach, practice) or the number of desensitizing verbs (e.g., touch, kiss) [184]. We also noted deep learning approaches that aimed to model the language in the text. Neural networks such as works by [91, 67, 146] used a bag-of-words representation that summarizes a conversation as the number of occurrences of each word in the vocabulary, regardless of the order in which they appear and do not require an explicit set of features.

User-based features, which are characteristics of a user's profile that can be used to make a judgement on the role played by the user in an online exchange and include age, gender, and sexual orientation, etc., were also common (N=20, 27%). Some papers considered time or/and location (N=14, 19%) as features. For example, Elzinga et al. [89] analyzed the change in the relationship between the offender and the victim over time to further examine how the threat level changed over the course of the conversation. Semantic features (N=12, 16%) represent the basic conceptual components of meaning for any lexical item. For instance WordNet [93] classifies words and uses hyponymy and hypernymy to establish semantic relationships between synsets that some papers such as the work by Iqbal et al. [136] used. In some papers (N=11, 15%), we saw the use of linguistic style or metrics of linguistic complexity that can help the model to distinguish language from the writing style of users. These metrics include Linguistic Inquiry and Word Count (LIWC)

(vocabulary lists known to help measure emotion in text) [220], readability, coherence, perplexity measures, and subjectivity measures. The keyword extraction method was used by 10 (14%) papers to extract keywords, it included named entity recognition, which involves extracting entities (names, location, email addresses), word clouds or tag clouds are another example of keyword extraction. Next, there were some other features that were used by the literature less frequently including, syntactic features (N=8, 11%), sentiment features (N=7, 10%), image features (N=5, 7%), network features (N=4, 5%), topic modeling (N=3, 4%), and relationships (N=3, 4%).

Behavioral features of predators or victims were also utilized (N=10, 14%); Vartapetiance et al. [279] used the Cycle of Entrapment (deceptive trust development, sexual grooming, isolation) from the Luring Communication Theory (LCT) [208] to differentiate predators by the process of entrapment used by child sexual predators to lure their victims into sexual relationships. Perpetrators usually approach the victim to build not only sexual but also emotional relationships [208]. Similarly, another study included as feature, fixated discourse, which is the unwillingness of the predator to step out of the sex-related conversation even if the potential victim wants to change the topic, as a feature [46]. These last set of features are a good example on how to leverage the knowledge from other fields that have studied the different stages and processes behind online sexual grooming to guide the feature engineering process. We observed that papers used a variety of features for detecting sexual risks and among them textual features were most popular.

Algorithmic Approaches: Mostly Traditional Supervised Machine Learning

We observed trends in terms of approaches used by sexual detection literature (demonstrated in Figure 2.3). Traditional or classic ML approaches were used in earlier studies in the area of pedophile detection by researchers. After Traditional ML approaches, system architectures and hand-crafted/rule-based models came to the picture. In the latest years deep learning models have

been used in the literature more recently since 2016, and in the last year an ensemble approach was proposed.

The majority of the studies relied on traditional or classic ML algorithms such as Support Vector Machines (SVM) (N=48, 66%) (Approach types by references are displayed in Table A.5 and algorithms' names are in Table A.6 in the appendix). For instance, Pendar et al.'s study [218] was one of the first works that used SVM and K-Nearest Neighbors (KNN) on the PJ dataset. Fauzi et al. [92] found that soft voting based ensemble for distinguishing predatory conversations from the normal ones performs the best on PAN-12 dataset compared to separately using Naive Bayes, SVM, Neural Network, Logistic Regression, Random Forest, KNN, and Decision Tree. That said, the study found that Naive Bayes outperformed other classifier models for differentiating between a predator and a victim in predatory conversations. Despite the increasing body of research that incorporates deep learning models, we were only able to identify a relatively small subset of the corpus that used deep learning in their approach (N=11, 15%). Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Long Short-term Memory (LSTM) methods have been used. Most of these models were for detecting sexual harassment, abuse, or sex trafficking (N=6, 8%) rather than sexual grooming detection (N=3, 4%). In one of the studies, Misra et al. [192] encoded predatory behavior purely from style (characters) to do authorship attribution on PJ corpus using a CNN model. One of the reasons that deep learning models were used less in the literature might be due to lack of large datasets since most deep learning models are known to require large amounts of training data. For instance, for sexual harassment detection using deep learning, researchers used publicly-available dataset SafeCity which includes 9,892 stories [144, 302, 171].

The papers that used ML or deep learning models mostly used supervised algorithms (N=48, 66%), 6 (8%) using unsupervised, with 4 (5%) using semi-supervised, and 4 (5%) using both supervised and unsupervised. Given the detection or prediction problem specifications, it is normal to find the

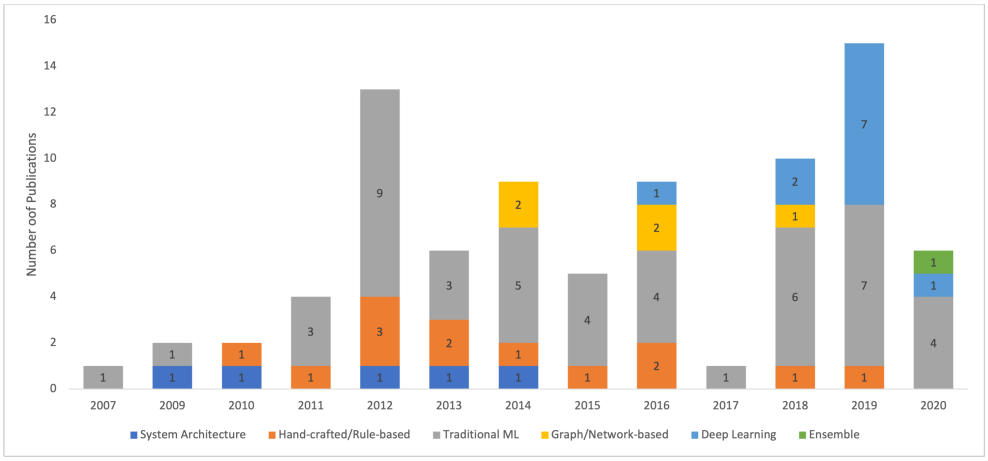

Figure 2.3: Frequency Distribution of Approaches in Reviewed Papers Over Time

supervised approach as the dominant selected type since in supervised approach, an input including the data labels is required for the models to be able to detect or predict the object of interest (users or types of conversations). We also noticed that quite a number of papers used semi-supervised approaches that only required a small amount of labeled data along with a huge dataset with a few number of labels as input for the models. Although a Semi-supervised approach can help address either the lack of available annotated datasets or the lack of annotators for a huge dataset, sometimes the models do not detect or predict accurately. For example, Kostakos et al. [158] used a semi-supervised learning approach with data collected from a popular European adult forum. Only 78 users were labeled based on their risk evaluation and left out 28,832 users for the model to predict their risk assessment, which their best model yielded only 79%. Unlike the two previous approaches, unsupervised approaches do not require annotated data as the models' input. With an unsupervised approach, models can capture patterns and extract knowledge from the input through algorithms, in most cases, clustering algorithms. For example, Toriumi et al. [271] was capable of uncovering risky communication behaviors on private chat systems used by minors to provide effective monitoring based on active communications.

Other than ML approaches hand-crafted or rule-based models were also commonly chosen as the methodology (N=13, 18%). For instance, Vartapetiance et al. [279] approach was built heuristically based on the Cycle of Entrapment which we discussed earlier. They created several categories for classifying conversations based on keywords presented and defined thresholds for assigning conversations to those classes. A few studies (N=5, 7%) used graph and network algorithms, in which all were detecting sex trafficking. Ibanez et al. [133] demonstrated that the content available in online escort advertisements can be used to identify provider networks and potentially roles using network graphs. We noted a lack of graph and network analysis for other types of risks such as sexual grooming or harassment on social media where considerable risks happen by analyzing social networks of users. A few papers proposed system architecture or tools (N=5, 7%) that were mostly based on filtering and blocking certain information that might cause risks for minors. Some (N=3, 4%) of these papers focused on sexual grooming [177, 219, 164]; for instance, [177] proposed architecture for detecting intent, time and location to block messages when children chat online. The other two papers focused on sex trafficking [288, 254]; for example, Silva et al. [254] proposed a system that filters and retrieves textual and image data related to sex trafficking to help law enforcement agents. On average, these systems were proposed in the year 2012 and they only block certain information such as address or phone numbers.

Task Granularity: Primarily Conversation Level Detection of Predators

Our results about the granularity level of the risk detection tasks for models on conversations (illustrated in the Table A.7) reveals that most of the studies that we reviewed focused on detecting and differentiating predator users and victims (N=27, 36%), rather than identifying patterns and changes that are indicators of risks. Some studies tried to identify predatory conversations (N=15, 20%). Kim et al. that [146] attempted to first classify each message and then used those results to classify the entire conversations using RNN to overcome the high potentiality of having

large blocks of conversations which might include hundreds of messages. Overall researchers were successful in the task of differentiating users with good algorithmic performance metrics. Some studies aimed to identify patterns (N=9, 12%) or characteristics/structures that contribute to being a predator in conversations. These patterns were usually associated with how predators approach victims and common methods used by them driven from theoretical frameworks for sexual grooming. Zambrano et al. [305] used existing theoretical frameworks to map different stages of the life cycle of the grooming (including gathering information, gaining access, lateral movement, escalating privileges, execution, debrief) within conversations of predators. They framed grooming as a vector of attacks which could be used for determining patterns of malicious behavior online. A few reviewed papers tried to identify predatory lines (N=8, 11%) in the conversations. Bours et al. [51] looked at individual messages to determine if such a message belongs to a sexual predator or not. While only two studies (3%) [238, 250] tried to identify the predatory risk level. Ringenberg et al. [238] used Fuzzy Sets for labeling messages for three levels of risks (low, medium, high), and developed a NN model that uses these fuzzy membership functions of each line in a chat as input and predicts the risk of interaction.

Algorithmic Output: Mostly Binary

Our analysis of the output types of algorithms is summarized in the Table A.8 in the Appendix, we list the publications grouped by the specific classification set up used. In there we show that most papers used a binary classification setting (N=44, 60%) where they usually classify conversations to predatory or not predatory, and users to predator or victims. Several studies performed multi-class classification (N=24, 33%) by differentiating class type rather than trying to fit all different types into binary classification. Khataua et al. [145] classified different types of sexual violence and associated risk of them to 4 different categories based on the locations that it occurs. Some papers (N=7, 10%) tried clustering methods, for instance, Kontostathis et al. [155] aimed to cluster

different types of predators via their language pattern usage using K-means on PJ dataset, and found 4 clusters produce the best results, meaning there exist four different types of predators. A few studies identified predatory stages (N=5, 7%), such as Potha et al. [224] where they examined the question set of each predator with a view to capturing the tone of predator's defensive or aggressive questions in order to identify patterns of predator's behavior that can be generalized in a real-life conversation as time series, i.e. using windows that only capture a short period of the predator's attacking strategy.

Overall, in this section we reviewed the models and the specification for these models. In the next section, we will report our findings about the artifacts that were built by the reviewed studies and will discuss how they evaluated the models.

Evaluation and Performance Metrics (RQ3)

In this section, we discuss our results regarding the evaluations of the performance of the artifacts produced from the reviewed studies.

Evaluation: Focus on Numerical ML Performance Metrics

For the evaluation of the performance of the models, most articles that we reviewed focused on computational performance (N=57, 87%) and did not leverage user studies or human evaluations to assess their approach. Only 3 papers (4%) [219, 164, 120] performed user studies to evaluate their models. Latapy et al. [164] evaluated their tool for automatic detection of paedophile queries by human experts that work in law-enforcement agencies and well-established NGOs. The reviewed articles mostly reported for Accuracy (N=39, 53%), F measure (N=27, 37%), Precision (N=17, 23%), Recall (N=14, 19%), and other numerical performance measures. Some papers only

measured accuracy [184, 114], which is not enough for assessing the performance of a classifier without other measures as it might be biased toward the majority class [207], especially in this application in which the dataset is imbalance toward non risky instances. PAN-12 was the first to benchmark performances of the models by the standard Information Retrieval measure of Precision (P), Recall (R), and F measure (weighted harmonic mean between Precision and Recall) [134]. Inches et al. [134] pointed out that the standard F measure equally weighted with precision and recall is not always desired. For the problem of identifying predators, detecting the right suspected users as predators is more important (precision) than having more suspects (recall). Since police officers would rather act fast towards the "right" suspect rather than "all" the possible suspects. Therefore, they used a measure of F with the factor equal to 0.5, to highlight precision. On the other hand, for the problem of identifying predatory lines in conversations, it is more important to retrieve more relevant lines (recall) to be used as strong evidence toward a suspect. So they used a measure of F with the factor equal to 3 to highlight recall. Overall, for the PAN-2012 competition, participants were able to detect the predators with accuracy of 93%. But for the second task of identifying predatory lines, researchers were not successful (47% accuracy) [134]. Researchers continued to improve these computational performances for these tasks even after PAN-2012. For instance, Kim et al.'s [146] recent study compared their results to the 16 competitors at the PAN2012 cyberpredator detection competition [134] and claimed that their results placed them first with respect to recall and F1 score, third with respect to F0.5 score, and fifth with respect to precision. However, they claimed that recall was the most important measure for the problem of detecting predators because it helps determine the fraction of predators who would go undetected, which is in contradiction of Inches et al.'s statement [134] that precision is the most important factor.

Application and System Artifacts (RQ4)

In this section, we discuss the artifacts and applications produced from the reviewed papers and usability of these algorithms and artifacts.

System Artifacts: Mostly Algorithms Only

An emerging theme from the literature in this area is their focus on the development and improvement of risk detection algorithms which resulted in improving the computational results of algorithms (N=63, 86%) rather than creating a system-based artifact. A few studies (N=6, 8%) [219, 250, 288, 254, 89, 155] created Forensic Investigation Tools to facilitate analysis for law enforcement to detect predators in chat conversations or find predatory relationships. Laorden et al. [163] created a conversational agent (Negobot) that posed as a child in chats to detect conversations with paedophiles. The aim of this research was to gather enough evidence from users suspected as paedophile to help the authorities to detect and identify paedophile behaviours. Michalopoulos et al. [190] created an Android application that detects sexual exploitation attacks by capturing incoming SMS messages, processing it with the assistance of classification and clustering techniques in a distributed topology. Additionally, there is a lack of APIs that could be integrated into social media, chat rooms, forums, and other platforms for detection of online sexual risks.

Intervention: Focused on Detection without Mitigation

Most of the literature did not suggest or implement intervention strategies (N=68, 93%). Only 5 (7%) papers [177, 219, 288**?** , 190] provided some form of intervention strategies. For instance, as previously mentioned Michalopoulos et al. [190] developed a sexual risk detection Android application, which provided an intervention method to send a warning signal to the designated

parent who is responsible for further actions in case of high risk. MacFarlance et al. [177] developed a prototype of an agent mediated autonomous system to automatically detect and block the transmission of personal data when children chat on-line, to detect and prevent attempts by users to arrange meetings. Most of the previous literature proposed systems for sexual risk detection without providing interventions for keeping users safe from online sexual risks.

Discussion

In this section, we provide a summary of key trends we found within the literature, discuss the implications of our findings, and reflect on potential gaps and opportunities that we found in our review of the literature. We also provide directions for future research in the area of sexual risk detection. Then, we reiterate the need for human-centeredness in computational risk detection.

Trends, Gaps, and Opportunities for Future Research on Sexual Risk Detection

Sexual Risk Detection is Skewed toward Sexual Grooming

We found a larger proportion of studies focused on sexual grooming, rather than other sexual risk types. The skewed direction of studies on sexual grooming detection might be due to the PAN 2012 conference competition, which led to a significant increase in publications in those years. In addition, there might be other underlying reasons for the decrease of literature in recent years, such as the iterative process that we used for finding relevant papers by cross references citations and the emergence of a novel coronavirus that was first reported in late December of 2019 [122]. Therefore, one suggestion to bolster more research towards tackling other types of sexual risk detection would be to hold hackathons or similar competitions to have groups of researchers converge on these topics. In more recent years, we saw an emergence of work on sex trafficking,

sexual abuse, and harassment, which we encourage the uptick in this trend. The long-lasting impacts on victims, as well as the distinctive characteristics of relationships that engender sexual risk (often someone close to the victim like a family member) [231] together underscore a necessity to close this gap in examining sexual abuse online. However, it would also be valuable to study factors that are pre-cursors to sexual violence, such as unwanted sexting requests. Sexting involves sending, receiving, or forwarding of sexually explicit messages, images, or media to others through electronic means and is prevalent among adults [278] and is becoming prevalent among youth [203, 231]. Sexting involves inherent risks given the possibility of negative outcomes involving bullying, non-consensual dissemination of sexual imagery, or increased violence against women [159]. Therefore, sexual risk detection for sexting and other types of unwanted (or wanted, in the case of minors) could be an important future direction for the sexual risk detection literature.

Datasets that Reflect Real-World Interactions and Users are Needed

It is crucial to train models utilizing real-world data that are representative of the target users, in order for the approach to be usable in the wild [39]. Yet, we found that most studies were based on public datasets. For instance, datasets analyzed for sexual harassment or abuse were based on public posts on social media, such as Twitter. Yet, we assert that analyzing public discourse about harassment and abuse is not enough for tackling this problem, as most sexual risks occur in private channels [303]. Thus, future works should leverage conversations in private channels for training their models. We understand that the data collection efforts associated with this recommendation are not trivial. It would involve researchers to engage in public scholarship and recruitment efforts to reach sexual harassment and abuse survivors, convincing them to share the details (i.e., digital trace data) of their most intimate and traumatic sexual experiences with researchers who they do not know or trust. As such, we encourage researchers to be thoughtful in how they engage in scholarship with vulnerable populations, in a way that respects their dignity and their privacy [287,

183]. Moreover, collecting private data by researchers and privacy constrain of not sharing the data publicly pose more challenges in terms of replication of research for the research community. However, a way to mitigate the concerns of replicability is to form coalitions of researchers who work together to solve important problems as a community of scholars who are committed to the protection of human subjects and data privacy. Another fruitful path of research would be to explore data anonymization approaches for ensuring the confidentially of participants when sharing datasets semi-publicly with other researchers through well-thought-out data sharing agreements and prior consent of research participants.

Further, the datasets reviewed in our paper were heavily skewed toward Perverted Justice (PJ). As the data acquisition for underage victims or law enforcement officers posing as children is challenging due to the laws and procedure involved [134], the PJ dataset focused on data from predators and adult volunteers posing as children. A limitation of this dataset is that it is not reflective of real children's interactions with predators, since the participants were adult volunteers. Approaches with datasets consisting of volunteers posing as children are potentially problematic, as it is possible to identify adults pretending to be children based on writing styles and authorship attributes [27]. While this approach may be adequate when the sole purpose is to detect predatory behavior, such an approach is not effective when the goal shifts to detecting behavioral patterns of youth victims. For instance, qualitative researchers found that adolescents struggle to handle sexual solicitation from people they know, but it is easier for them to reject sexual solicitations from strangers [231]. Therefore, taking a complementary approach of detecting strangers combined with detecting potential victim-signaling behaviors of teens could potentially bolster efforts to intervene and advocate for youth before victimization occurs. To do this, however, future research needs to create datasets that are ecologically valid for youth and use conversations from real adolescents to train such models.

Establishing Ecologically Valid and Trauma-Informed Ground Truth

Risk is a highly subjective construct, and defining and quantifiably operationalizing risk has substantial impacts on risk management and safety [222]. Through our literature review, we found that what denotes and marks sexual risk behaviors online has not been grounded in a systematic way. In most cases, literature in our review did not report IRR to establish consistency of their ground truth annotations. Although theory was sometimes used for feature selection, such grounding was not used for data annotations. Theoretical design [35] utilizes a wealth of concepts and theories from behavioral and social sciences toward creating high-quality ground truth annotated datasets. For instance, future studies could leverage existing theoretical frameworks related to online sexual risks for establishing ground truth for data annotation. Computational scientists should form more collaboration with social scientists to build more frameworks for data ground truth for emerging sexual risk issues. The subjectivity of risk, in these cases, negatively impacts establishing robust ground truth [160] for sexual risk detection datasets used for ML risk classification. Researchers need to fill this knowledge gap by establishing an understanding *what* the most salient online sexual risk factors are through social science and psychological theories and by engaging directly with clinical experts. Given the highly subjective nature of risk, it is important that studies provide clear explanations on how they defined sexual risks for ground truth annotations, annotation procedures, and the annotators' backgrounds. It is essential to clearly explain the data annotation process, include the definitions that were used for data labels, consider annotators' demographics and expertise, and ensure sufficient inter-rater reliability, to address any potential bias from human annotators. These labels should be informative, discriminating, and independent. In data labeling, domain knowledge and contextual understanding help annotators to create high-quality datasets.

Including clinicians and subject matter experts in the annotation process would help improve the overall quality of ground truth. People have different perspectives on sexual risks based on their

personal experiences and backgrounds; thus, it is important to look into annotators' backgrounds and experiences. Subject matter experts, such as clinical psychologists or those who have background and training in sexual trauma, would be more equipped to annotate these datasets [11] than research faculty and/or their students. In addition, people who have personally experienced some form of sexual risk or abuse may have a better understanding of these types of risks than people who have not, making sexual abuse survivors another group of people suitable for the annotations. Yet, we make this recommendation with caution; asking past sexual victims to annotate sexual risks could inflict trauma by triggering memories of the annotator's own abuse experiences [257]. Therefore annotation process should be completed with utmost care and different techniques such as stress management, time management, relaxation, leisure, and personal renewal [304] with frequent mental-health checks should be used to make sure the annotators with past trauma are not experiencing any difficulties and are willing to continue the annotation. Since the current literature relies heavily on third-party annotations, they might not have the same perspectives as people who were actual victims of online risks [148]; future studies should also take into account the perspectives of the victims. There are ways of taking into account the perspective of victims without having them directly annotate their data, for instance, Kim et. al's work [148] where they used user-labeled posts on a mental health peer support to indicate ground truth for bullying experiences. Yet, we acknowledge that resource constraints and time limitations are reasons why researchers relied heavily on non-expert annotators. One possible middle-ground to overcome these limitations is to have experts create the codebook and guide non-experts performing the annotations.

Models Need to Consider Contextual Information

We found that most of the literature disregards the context of the risky interaction, which can provide important information for defining risks. For instance, most approaches focused on a single data type such as text, in the absence of images or available meta-data. Due to challenges

of multi-modal ML, such as representation, translation from one modality to another, alignment, fusion, and co-learning, most of the literature only focused on one modality [31]. Multi-modal data provides more context, especially in social media and online conversations. AI methods need to be able to interpret multi-modal signals together to make improvements in understanding the complexity of real-world experiences. Failing to include contextual data in the training of detection models is likely to result in a significant amount of false positives. A high false positive rate may cause systems to over-flag content, consequently reducing the system's capacity for effective risk mitigation. Although there are challenges in multi-modal approaches, improving the ability of models to process additional context can be a fruitful undertaking.

As many societal and psychological factors are at play when sexual risks happen, exploring mechanisms to model these factors in automated approaches becomes important, and we realize that there have been limited efforts in this direction. Although we observed good practices from the sexual grooming detection literature [279, 46, 208] on how to leverage knowledge and theories from other fields that have studied the different stages and processes behind online sexual grooming to guide the feature engineering process, these features have not been used as widely. Therefore, future works need to consider using these behavioral and contextual features. In fact, it would be interesting to explore if more recent approaches on representation learning are able to learn this type of latent information from the data. Further, some literature did not provide justifications for the features used. Although for some applications, the initial solutions for feature selection are often intuitive and based on human observation, there is still a need to perform post-hoc analyses of features to show their usefulness as well as how they connect with established theories. For grounding future computational models in human theories and knowledge, more data needs to be empirically analyzed by experts from humanities and social science fields to create new theories or frameworks that can be leveraged in AI models. For example, by analyzing social media data that involves risk, social scientists could create frameworks related to the ways in which youth may

unintendedly indicate their vulnerability to become more susceptible to such risks. Thus, these frameworks could be integrated as features, inputs, or design of algorithms and models for detecting such instances. Since in some cases the risks phenomenon that happen online is so new that theories might not yet exist to inform technical development.

Advancing the State-of-the-Art in ML through Deep Learning

We found that most approaches are based on traditional ML algorithms rather than deep learning algorithms; this might be partially due to lack of data, since deep learning models require large amounts of training data. Moreover, massive computing resources, and significant amounts of time are needed to successfully train deep learning models [66], and both of these items are often only available to big tech companies, and a handful of large research labs. As ML models are used with the final intention of supporting decision making for users, a great deal of information is needed in order to relate the user's decision to the solution given by the model. Deep learning models have shown promising performances in many tasks and domains when given a large enough dataset [66]. Therefore, we recommend collecting sufficiently large datasets to facilitate the use of deep learning models that could benefit risk detection systems in future studies. To overcome some of these data limitations, pre-trained deep learning models can also be used to address the lack of large labeled data which constrained some papers [85, 262, 251]. That said, given recent findings on the underlying biases that characterize pre-trained deep learning models, such as language models [105, 58], we suggest researchers to adopt ample caution when they are appropriated, such as considering debiasing approaches in concert [48]. Additionally, it could be interesting to explore data augmentation techniques to automatically supplement training data.

Another common shortcoming we observed is related to the classification setting of the task. The majority of the papers focused on binary classification, where the study would aim to identify

an online risk instance or a predator. Binary classification fails to fully depict the characteristics of different sexual risks. Each sexual risk has its own set of stages and levels, which are hard to differentiate under the lens of binary classification. Future research should gear towards multi-class classification based on the patterns of communication, focusing on different types of sexual risks as well as different stages and levels of each risk type. Multi-class classification will help researchers better understand sexual risks with respect to both categorical (types, levels) and temporal (stages) dimensions, which will be helpful in identifying online risk instances during the early stages, before the victim suffers from risk exposure for a prolonged period of time. Understanding the granularity level of analysis for conversations is important since approaches that focus on single messages would fail to understand the semantics and the context of conversations. Understanding individual messages is important when the context of the entire interaction is understood. Overall our results demonstrated that most approaches focused on identifying users, so future researchers need to develop approaches that take into account patterns and changes at the conversation level to detect the risks in earlier stages when it is happening.

Our finding that most studies used supervised algorithms implies that future works should develop more semi-supervised or a combination of supervised or unsupervised methods to provide automatic learning improvements through computational feedback. Particularly, unsupervised learning can address the challenges of not having big datasets and help provide insights about online risk. Unsupervised models can be considered as a first step used usually for understanding the underlying knowledge of the data to aid developers create supervised models based on accurate understanding of the data. The prior understanding of the data is important not only for models' accuracy but also for models' interpretation by humans. In addition, human-in-the-loop approaches and active learning (humans handling low confidence units and feeding those back into the model) [128] should be utilized so the models select what they need to learn next; that data may be sent to human annotators for training to teach edge cases, identify new categories, to help avoid over-fitting, or to

help adapt to changing data characteristics and contexts.

Objective Performance Evaluation Are Not Enough

A common theme among the sexual detection approaches was that they came from a purely computational perspective and failed to incorporate any aspects of user-centered design or needs analysis. The articles did not include formative or summative user evaluations of the solutions developed. Instead, the majority of the papers reported computational evaluation metrics to demonstrate the performance of their risk detection models. The difference in classification tasks and the variety of evaluation metrics used in each study make it difficult to develop a standardized benchmark for computational evaluation. Standard metrics for measuring algorithmic performance for each specific risk detection task and guidelines for evaluating solutions' effectiveness and performance are needed for a direct comparison of different algorithms for each task. Although, the current computational evaluation metrics are effective in the sense that they show computational performance, they could be further strengthened by assessments through user case studies and error analysis with examples. In addition to the algorithm evaluation metrics, it can be valuable to conduct user studies to evaluate the feasibility of the models in real life scenarios. Tests and validations that involve humans should not only be done by the researchers but also by the actual stakeholders of the system for a more accurate evaluation [261].

Researchers have proposed different categories for evaluation methods or scales that could be adopted to ensure proper evaluation of the systems. For instance, Mohseni et al. [200] divided diverse types of evaluation methods into two groups including objective (task performance, user prediction of model output, compliance/reliance, etc.) and subjective measurements (interviews, surveys, self-reports, etc.) to be performed for evaluating the systems. They also categorized evaluation measures in the five themes Mental Models (how the AI works), Usefulness and Satisfaction,

User Trust and Reliance, Human-AI Task Performance, and Computational Measures (correctness and completeness in terms of explaining what the model has learned). There exist some scales that can also be used by risk detection systems to evaluate their AI system, for instance the "Explanation Satisfaction Scale" by Hoffman et al. [127] consisting of 8 items, which address factors such as understanding, satisfaction, completeness, accuracy, or trust. In addition, participatory design [35] could be used to involve people in the design of the algorithms to address the gaps between technical solutions and human expectations. Therefore, using different evaluation methods and scales as a checklist of items to be evaluated should be employed by researchers and reviewed in surveys.

Need for Artifacts and Embedding Models in Real-World Systems

We encourage future research to go beyond creating algorithms to developing technologies similar to how they would be used in practice. The majority of the studies fell short of implementing the algorithms in real-world systems or applications. Designs for these systems and interventions should be characterized to the relevant social groups and stakeholders such as adolescents, parents, police investigation teams, online moderators, social media companies, researchers, policymakers, practitioners, etc. from a Social Construction of Technology (SCOT) perspective which advocates that technology does not determine human action, but that rather, human action shapes technology [41]. Bringing user-centered perspectives to the forefront of sexual risk detection is necessary to evaluate the system by involving targeted users with the targeted context to make sure the artifacts meet the needs of different social user groups and the characteristics of the respective stakeholders. When creating ML algorithms, testing, and tuning the dataset is also important. It is important that future work leverage humans in the various circles of creating ML models including training, tuning, and testing algorithms. At the same time, we do acknowledge that sometimes the choice of the ML model may limit the extent and utility of human involvement. For instance, deep learn-

ing models are inherently opaque and therefore model decisions or outcomes may not be easily understandable to laypersons in an intuitive manner. To support human involvement in such cases, researchers can consider approaches like "model cards" [197] to communicate to users about the limitations of the models and what the model outputs may mean.

From a practical standpoint, most articles that we reviewed did not declare the memory and the processing time of their approach, nor discuss the feasibility of implementing their model on mobile devices with fewer resources. Essentially as mobile devices are prevalent among youth which have their own specifications and limitations such as less computational power [264, 266]. Therefore to address the gap in the literature, researchers and practitioners need to design, implement and test approaches that work with mobile devices that considers computation, memory, and other limitations. Additionally, future researchers need to develop APIs that could be integrated into social media, chat rooms, forums, and other platforms for the detection of online sexual risks.

Furthermore, as we stated the gap of designing systems for respective stakeholders there are more points to consider when developing artifacts for different social user groups. These systems need to provide explainability of the AI system's functioning or decisions to be understandable by respected stakeholders [8]. These systems need to generate post-hoc explanations to justify an opaque model's decision in an accessible manner. Reasons for explanations include trustworthiness, causality, transferability, informativeness, fairness, accessibility, interactivity, or privacy awareness [308]. As the motivational, social, cognitive, along with professional and educational profiles of stakeholders affect their interpretations and reactions to explanations, so these explanations need to be suitable for each specific social group [8]. It is also important to explain why the stakeholders should trust the system. For instance, if the stakeholders are government agents or police officers, should certify the model compliance with the rules in force. Trustworthiness might be considered as the confidence of whether a model will act as intended when facing a given problem. There are cases in which the lack of a proper understanding of the model might drive

the user toward incorrect assumptions and negative consequences. In this case, for instance, police officers or parents need to know exactly how much they can rely on the model, by receiving flagged content on what it means and how much they can trust the system. For example, a teen could send a picture with swimsuits that is detected as a sexual risk, but the parents should know how false positives might happen and know when to trust the detection system and when to manually check. These models should assess a generalization of robustness and stability, and confidence so humans can make the decisions on how to trust these systems.

The Importance of Human-Centeredness in Computational Risk Detection

Toward Victim and Survivor-Centered Sexual Risk Detection

To date, the primary focus has been detecting online sexual predators to help law enforcement agencies identify and prosecute sex offenders [42]. Although identifying predators is an important task, future work should consider identifying behaviors or indicators of being a victim of sexual risk. Victims are equally, if not the most, important stakeholders in the sexual risk incidents; examining how we can identify them could lead to strategies to support victims and even prevent any potential victims from going through such traumatic incidents. Identifying victims' reports can help concerned parties to pay attention to the violence reports and take reactions in a timely manner. Thus, the timing of the risk detection is of critical important; the earlier we can detect risk exposure, the sooner we can intervene and mitigate the risks. Thus, artifacts and systems are needed to help prevent victimization rather than after-the-fact risk detection. Also, risk detection and mitigation come jointly; without a plan to mitigate risks, it would be useless to detect it. Therefore, when reviewing literature in the area of risk detection, researchers need to take extra factors into consideration, for instance, the nature of the risk and the context in which it happens, the break down of the computational results and how that may affect the end-users, the timing of the

detection, and the mitigation plan to protect the users. As HCI, HCML, and ML researchers, we are uniquely positioned, and thus arguably ethically obligated, to use our multi-disciplinary skills to not only accurately detect when sexual violence occurs online for the purpose of understanding this modern-day phenomenon, but to also become activists that work to eradicate sexually-motivated online violence.

Toward Human-Centeredness in Computational Risk Detection

Without applying the human-centered lens to the context of sexual risk detection, many of our insights would have been overlooked. We came to these insights about the gaps and opportunities in the literature by having the target users' needs in the valid real-world scenarios in mind. Many authors of the papers we reviewed did not mention the limitations of their research that we surfaced in this research. Some of these insights include the after-the-fact risk detection, lack of ecologically valid datasets, and lack of user studies and real-world evaluations. Overall, there were many valuable contributions made by the existing literature that should be continued in future work. For example, researchers from social sciences [33, 211] proposed theoretical frameworks for online grooming processes to help the developers of predators automated detection systems, leveraging the behavioral pattern recognition to improve educational tools for the community. For instance, Cycle of Entrapment from the Luring Communication Theory (LCT) [208] was developed by social scientists related to the different stages of how sexual grooming happens. This theory has/can be used to refine the data annotation and feature selection processes for AI model development to improve accuracy through insight into human behavior. This would contrast with solely data-driven approaches. Human knowledge also includes understanding social and human biases and trying to adjust that for algorithms to balance unfair decisions. The increase in human theory and involvement increases the interpretability and the practicality of the models, as these models, in the end, serve the sole purpose to aid and help the people. Therefore, we need increased but also

well-monitored and well-reformed involvement of humans in AI risk detection design. Similarly, we urge computational ML experts to work with HCI researchers and social scientists when dealing with real-world human problems to incorporate human and social interpretations in the design and development of the models. If we want to protect people from online risks we would need a close collaboration with policymakers, psychologists, and lawmakers.

Computational online risk detection, in the end, is a problem that originates between the online interactions between people and aims to protect people and mitigate any potential risks. Thus, our human-centered lens can also be applicable to other domains of online risk detection, such as cyberbullying or hate speech. Therefore, other researchers could use our human-centered lens for reviewing the computational risk detection literature beyond that of online sexual risk detection. Moreover, it is important that researchers and practitioners respond to the call for the need to add FATE (fairness, accountability, transparency, and Ethics) from the point of view of a human observer to explain the reasons behind decisions or the processes that generate them [26] which is missing from the current sexual risk detection literature.

Moving Beyond Human-Centeredness to Examine and Embed Values in HCML

According to Adadi et al. [8] the need for explaining AI systems comes from four motivations. First, there is the need to ensure that AI-based decisions were not made erroneously with justifications for each outcome. Second, explanations make vulnerabilities and flaws visible so that they could be controlled and corrected. Third, explainability is needed so others can continuously improve the models with a better understanding of the capabilities and implications of the models. Fourth, providing explanations could potentially lead to discovering new facts and gaining more knowledge. It is important to consider AI systems' fairness to avoid "unfair" algorithms/systems whose decisions are skewed toward a particular group of people. According to Mehrabi et al.'s

[187] survey of different biases in ML application, there are two sources of unfairness including those coming from biases in data and those stemming from algorithms. ML researchers need to use such taxonomies for fairness to avoid biases in the technologies that they develop.

Therefore, researchers need to think more about the ethics and speculate more on how these models could be used by bad actors. There are ways that sexual risk detection models could be used in adverse ways to harm people instead of helping them. As an example, sexual predators could use trained computer vision algorithms to find child pornography. So it is important to make sure that proper safeguards are in place to protect people. In order to do so, we advocate that studies can take a speculative approach [35] on how these algorithms might be used to harm. Moreover, privacy awareness of the model by users is important since models may have complex representations of their learned patterns and not being able to understand what has been captured by the model and stored in its internal representation may entail a privacy breach. Thus for researchers and practitioners that develop these types of systems need to think about potential privacy breaches of users and potential misuses of the ML model. There was a universal lack of any explicit mention of the ethics, potential biases or speculations on the usages of the models in the reviewed papers. Exceptions include Chowdhury et al. [108], who discussed privacy of individuals, fairness, bias, discrimination, and interpretation of their study. The literature being scant on FATE topics was our barrier to qualitatively coding the papers for these values in our literature review in a systematic way. Future studies should move beyond considering only computational approaches and consider more human values embedded in their research, following the recommendation of Jo and Gebru [141] around requiring and building suitable institutional frameworks and procedures.

Conclusion

We reviewed 73 studies on computational approaches to online sexual risk detection utilizing a human-centered lens, which may be used by researchers for reviewing computational risk detection research more broadly. Our review provided insights into the trends and gaps within the current literature and opportunities for future research. Although our literature review is on computational methods, as sexual risk detection is a social issue that needs integration of personal, social, and cultural aspects for designing and developing intelligent systems that understand this socially complicated issue, reviewing literature from a human perspective is necessary. We call on the community to design and develop human-centered solutions to address these gaps by considering the stakeholders at all stages of designing and developing technologies that bridge the socio-technical gaps for sexual risk detection.

CHAPTER 3: STUDY 1: LET'S TALK ABOUT SEXT: HOW ADOLESCENTS SEEK SUPPORT AND ADVICE ABOUT THEIR ONLINE SEXUAL EXPERIENCES

Citation: A. Razi, K. Badillo-Urquiola, P. Wisniewski, "Let's Talk about Sex: How Teens Seek Support and Advice about Their Online Sexual Experiences", 2020 ACM CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA. https://doi.org/10.1145/3323994.3372138 (23% acceptance rate)

Based on our literature review findings, we realized that an empirical analysis of teens' online sexual experiences as needed to development human-centered knowledge to inform the design of our sexual risk detection algorithms. Therefore, we conducted a thematic content analysis of 4,180 posts by adolescents (ages 12-17) on an online peer support mental health forum to understand what and how adolescents talk about their online sexual interactions. Youth used the platform to seek support (83%), connect with others (15%), and give advice (5%) about sexting, their sexual orientation, sexual abuse, and explicit content. Females often received unwanted nudes from strangers and struggled with how to turn down sexting requests from people they knew. Meanwhile, others who sought support complained that they received unwanted sexual solicitations while doing so—to the point that adolescents gave advice to one another on which users to stay away from. Our research provides insight into the online sexual experiences of adolescents and how they seek support around these issues. We discuss how to design peer-based social media platforms to support the well-being and safety of youth.

Introduction

Youth leverage the internet to seek support and advice about relationships [147] and their sexual health [143, 263]. Yet, less is known about how they seek advice and support via the internet about their online sexual experiences (i.e., sexual interactions that are mediated by technology). In 2014, Weinstein et al. [291] found that using the internet as a means to seek intimacy with others has become a key stressor for adolescents. These researchers were among the first to analyze digital trace data from adolescents to gain deeper insights into their online experiences. We build upon this prior research by analyzing social media trace data from an online peer support platform to understand how adolescents seek support and advice regarding their online sexual experiences. We pose the following high-level research questions:

- RQ1: For what purpose do adolescents post on online peer support platforms when discussing their online sexual experiences?
- RQ2: a) What types of sexual experiences do adolescents most frequently post about? b) What are key characteristics of these experiences?
- RQ3: What are the consequences of adolescents experiencing these online sexual interactions?
- RQ4: What are the key challenges associated with adolescents seeking online peer support about their online sexual experiences?

To answer these questions, we conducted a thematic analysis of 4,180 posts made by 3,034 adolescents (ages 12-17) on an online peer support platform for youth and young adults. We found that adolescents used the platform to seek support, connect with others, and give advice about their online sexual experiences. For the most part, females were the primary posters, asking for advice

on how to combat peer pressure when a romantic interest or friend asked for nudes. A significant number of adolescents expressed concerns about unwanted sexual advances, while they were seeking support online—to the point that some posts warned adolescents about other malicious users. Another major theme was that adolescents tried to connect with others who had similar life experiences as them. Our study makes the following research contributions:

- Analyzes social media trace data to provide an unfiltered view into the lived online sexual experiences of over 3,000 adolescents.
- Provides a deeper understanding of peer support seeking behaviors for their online sexual experiences.
- Uncovers the challenges of online peer support and makes recommendations for design.

Our research makes empirical contributions to the fields of Human-Computer Interaction (HCI), adolescent online safety, and online peer support. Our findings will also help designers create online peer support forums that consider the vulnerability and online safety of adolescents.

Background

The HCI community has a history of engaging in important work related to sensitive online disclosures and sexual abuse [19, 21, 176, 182]. For instance, Andalibi et al. [19] investigated self-disclosures by sexual abuse survivors and the relationships between anonymity, disclosure, and support seeking. Similarly, our work studies self-disclosures about sexual abuse but more broadly examines other online sexual experiences. In contrast, our research specifically focuses on adolescents, who are developmentally different than adults. Below, we situate our research at the intersection of technology, sexuality, and online risks for adolescents. We highlight the gaps within the

literature to emphasize the contributions of our research.

Adolescents, Technology, and Sexual Exploration

Emergent technologies, such as social media and mobile smartphones, have provided new means for adolescents to explore, seek advice, and talk about sensitive topics, such as their sexuality, sexual health, and sexual experiences [98, 263, 292]. For example, Suzuki and Calzo [263] studied two online bulletin boards to understand the types of sexuality and relationship questions adolescents asked their peers. They found that adolescents often posted about their physical, emotional, and social selves. Romantic relationships and sexual health were the most common topics discussed. In contrast, Forte et al. [98] surveyed high school students about their social media and information-seeking behaviors. They found that students frequently asked their social networks about school-related topics, rather than sexual identity or health. While this work demonstrates that adolescents leverage the internet to seek support and information, it also suggests that adolescents use online platforms differently based on the context (e.g., health) and audience (e.g., classmates). Further, this research gives us insights into how adolescents seek support for sexual health in general, rather than technology-mediated sexual experiences. We build upon this literature by examining how adolescents seek support about online sexual experiences via online platform for peer support.

Adolescent Online Safety and Sexual Risks

The bulk of research around adolescents and their online sexual experiences has primarily focused on sexual risks, including sexting (e.g., [81, 71, 157, 193, 239, 276]), unwanted online sexual solicitations [196], and online sexual grooming [103]. In 2018, Sklenarova et al. [255] conducted a survey with German adolescents to understand their experiences with unwanted sexual solicitations.

They found that 51% of adolescents had experienced online sexual interactions, mostly with peers. Only 10% of these experiences were perceived as negative or unwanted, but those who lacked social support received more unwanted solicitations [255]. While such findings are valuable, most of our knowledge about what adolescents are doing online, as well as the outcomes associated with these activities, is derived from large-scale surveys [142, 165] or smaller-scale interview studies [52, 180] that rely heavily on self-reports [223], which may be limited by recall bias [111] and social desirability bias [96]. Only a few studies have leveraged digital trace data of adolescents. In 2014, Doornwaard et al. [80] friended 104 Dutch adolescents on Facebook and conducted a content analysis of adolescents' friend-based interactions and uncovered that about a quarter of the profiles contained sexual and romantic references. A clear theme from the literature is that adolescents frequently experience sexual interactions online; yet, there is still more to uncover in terms of how adolescents seek support and advice regarding these online sexual interactions. Further, more research needs to examine the interplay between adolescent mental health, online peer support, and the online sexual experiences adolescents discuss via such platforms [290]. The contributions of our work lie at these intersections.

Methods

We analyzed digital trace data from an online peer support platform. The platform operates primarily as a mobile app targeted towards youth to provide mental health support. The primary researcher licensed the dataset that contained about five million posts and 15 million comments made by approximately 400 thousand users. The dates of the posts ranged from 2011 to 2017. We anonymized the name of the platform to protect the identity of the adolescents included in our data analysis. Approximately 70% of the users on this platform are between the ages of 15-24 with 6% (21K users) between the ages of 12-15. We filtered posts by age to only include adolescents

between the ages of 12-17. While nationality was not a variable in the dataset, most of these youth were English speakers primarily from the United States, United Kingdom, and Europe.

Considerations for Data Ethics

Our Institutional Review Board (IRB) determined that this research was exempt from human subjects' review because personally identifiable information (e.g., usernames) was removed prior to sharing the dataset with the researchers. However, we still took the utmost care to preserve the confidentiality and privacy of the adolescents within our dataset. Due to the complex nature of the dataset (e.g., open-ended text responses without images), we took an extra precaution of requiring all research assistants who had access to the dataset to complete IRB CITI Training for working with human subjects before working on this project. We also made sure that any personally identifiable information was removed prior to including any quotations within our results. We confirmed that none of the quotations included in this paper could be tied back to the original poster through a search on Google. Additionally, due to the explicit nature of many of the posts (e.g., describing sexual assault, sexual acts, self-harming behaviors, etc.), we took special care to continually assess the mental health vulnerabilities of our research team [16]. As a result, we revised our scoping criteria (as discussed below) to remove self-harm content that was less relevant to this analysis and reassigned one undergraduate research assistant to work on another project that was less triggering for them. There were no cases where we had access to personally identifiable information for individuals reporting sexual abuse or imminent risk to a minor. Therefore, we did not need to invoke our status as mandated child abuse reporters during this research project.

Scoping and Relevancy Coding Process

Given the dataset included over five million posts, the goal of our scoping process was to scale down the dataset to a practical size for qualitative analysis. Therefore, we focused our investigation on original posts (not comments) made by adolescents (ages 12-17) that discussed online sexual experiences, behaviors, or interactions. To do this, we first included search terms for popular social media platforms between the years 2011 to 2017 [2], which included Instagram, Kik, and others. Then, we searched for common sexual jargon used by adolescents [268]. Next, we supplemented these terms through an exploration of the dataset. research assistants read over 5,000 posts to generate a list of keywords relevant to our topic. Once we reached a saturation point where reading an additional 50 posts did not result in additional keywords, we concluded this process. The final lists of keywords (Table 1) were grouped conceptually at the intersection of "online" terms and "sexual" terms. The first author created a SQL query to identify posts that contained these terms, and the research team worked together to refine and optimize the query for relevance. The final data query resulted in a record set of 8,271 posts made by 6,351 adolescents. The posts were then divided among five research assistants for relevancy coding. Two independent coders read each post and coded it based on the following inclusion criteria:

- 1. The post involved an online component—beyond that adolescents were posting, and
- 2. The post discussed sexual topics, such as sexuality, sexual behaviors, and/or sexual experiences.

For example, the post below came up in our search results (based on the words in bold):

"I smashed my iPod by accident and haven't been able to get a new one until now. Nothing serious.... I'm okay :') I'm hear for anyone who needs to talk to me... At ANYTIME!! <3 Kik: [Kik ID]" 14yr old Female

Table 3.1: Scoping Search Terms

	Keywords
Online Terms	Facebook, Instagram, Tinder, Bumble, Grinder, Snapchat, Craigslist, Skype, Hinge, Whatsapp, Kik, Discord, Messenger, Omegle, Vimeo, Vine, Tumblr, Myspace, 4chan, Reddit, forum,blog, video chat, Facetime, ft, message, dm, sent, send, pm, online, meet on, met on, webcam, gaming,cyber, blackmail, internet, AMOSC, f2f, LMIRL
Sexual Terms	Sex, nude, naked, flirt, STI, STD, grooming, LDR, predator, rape, solicit, dick,threesome, 3some, pussy, vagina, penis, cock, cunt, anal, clit, clitros, thick, boob, breast, tit,nipple, oral, sodomy, finger, handjob, touch, balls, fondle, birth control, BCP, plan b, condom,#metoo, nonconsensual, pedophile, catfish, BDSM, bondage, dominant, sadism, masochism, lesbian, gay, cougar, smash, virgin, underage, minor, nsfw, make out, made out, sugarbaby,horny, LEWD, blowjob, BJ, friends with benefits, DFT, hentai, porn, dry hump, Netflix and chill, thirsty, TDTM, cum, sperm, semen, cunnilingus, dildo, ejaculate, masturbate,erect, fellatio, foreplay, foreskin, genital, hepatitis, herpes, homo, hymen, IUD, lube, morning after,morning wood, libido, hickey, lick, one night stand, orgasm, rimming, scrotum, vibrator.

The post was deemed irrelevant because it had an online component but did not discuss anything sexual (the term "smashed" was not used in a sexual manner). Posts were removed because they were surveys (26%), duplicates (5%), public service announcements (e.g., phone numbers to hotlines) (4%), unoriginal content (e.g., song lyrics) (3%), or were not written in English (<1%). We calculated Interrater Reliability [259] on our relevancy coding and found substantial agreement (Cohen's kappa=0.71). To resolve conflicts, we formed a consensus among all five coders. Our final dataset included 4,180 relevant posts.

Data Analysis Approach

We conducted a qualitative thematic analysis [53] of the adolescents' posts. First, two independent raters coded 10% (N=418) of the posts to generate initial codes for the analysis. Five research assistants split up these posts and met to form a consensus across their codes and create a master codebook. Then, two independent coders (of the same group of research assistants) recoded this data to confirm IRR, which ranged from substantial agreement (0.71) [259] to a complete agreement (1.00). Next, the remainder of the dataset was divided among the five coders for coding. Once the data coding was complete, the first author reviewed the codes and used an axial coding process [70] to merge similar codes, groups codes conceptually by theme, and identify emerging patterns. Table 3.2 presents the final codebook dimensions, themes, and codes. To address RQ1, we identified three primary types of posts: 1) Seeking support from others, 2) Trying to connect with others, and 3) Giving support to others. Across all three types of posts, we identified the following types of sexual experiences (RQ2a) being discussed:

Table 3.2: Final Codebook Dimensions, Themes, and Codes

Purpose of Posts (RQ1)	**Seeking Support (83%, N=3,474)**	**Trying to Connect (16%, N= 635)**	**Giving Advice (5%, N=200)**
Types of Teen Online Sexual Experiences(RQ2a)	Sexting(78%, N=2706) Sexual Orientation(16%, N=549) Sexual Abuse(8%, N=292) Explicit content(7%,N=237)	Sexual Orientation(62%, N=392) Sexting(37%,N=234) Sexual Abuse(5%, N=30)	Sexting(62%, N=123) Sexual Orientation(27%, N=54) Sexual Abuse (21%, N=42)
Characteristics of Online Sexual Experiences(RQ2b)	**Situational Context**	Codes: Initiator, Recipient (unwanted)	
	Relationship Context	Codes: Stranger,Acquaintance/Friend, Dating, Family	
	Copying Response	Codes: Engaged, Rejected(Blocked or Reported), Did nothing	
Consequences Of Online Sexual Experiences(RQ3)	**Consequences**	Codes: Mental Health, Bullying, Exposure, Blackmail, Positive Feelings	

- **Sexting:** Posts that mention sending or receiving sexually explicit messages or photos, cybersex, or other sexual exchanges online.
- **Sexual orientation:** Posts exploring one's gender identity or sexual orientation.

- **Sexual abuse:** Posts discussing a sexual violation, harassment, abuse, or aggressive sexual behavior
- **Explicit content:** Posts about viewing sexual content online, such as pornography.

Finally, for the posts that sought support about online sexual experiences (N=3,474), we identified the following salient characteristics of the interaction **(RQ2b)**:

- **Relationship context:** The relationship (e.g., stranger, acquaintance, or dating) between the adolescent and person in which they had the online sexual experience.
- **Situational context:**Whether they initiated or were the recipient of an unwanted solicitation
- **Coping Response:** How they responded (e.g., engaged in the activity, rejected an advance, or did nothing).

Finally, we analyzed the consequences (RQ3) associated with each type of sexual experiences, which included mental health related problems, bullying or harassment, unwanted exposure, blackmail, or positive feelings. We allowed for double-coding and did not apply codes if they could not be determined from the post. Therefore, percentages in Table 2 may add up to slightly more or less than 100% for each dimension coded. We present our findings in descending order based on the frequency of the coded dimensions within our dataset.

Results

In this section, we present our results. We use 3.2 as the over-arching structure for this section.

Descriptive Characteristics of Adolescent Users

The 4,180 posts in our analysis were made by 3,034 unique adolescents. The adolescents were between 12 and 17 years old, with the average age (at the time of posting) being 15 years old. Most of the adolescents were 16 years old (28.2%), while the rest were 17 (26.3%), 15 (22.3%), 14 (15.5%), 13 (6.4%), and 12 (1.3%). Most of the posts were from female users (73%), with 16% from males, and 11% from non-binary or unspecified gender individuals. About 38% of the posts were posted anonymously. These adolescents had been active for an average of 7.5 months from the date they posted in the dataset. On average, the adolescents posted 206 original posts (SD = 436) and 898 comments (SD = 2,040). About a quarter of the posts (26%) specifically mentioned using other social media platforms. Of these posts, the majority mentioned Kik (43%), followed by the peer support platform (15%), Snapchat (12%), Instagram (8%), Facebook (8%), Skype (4%), Tumblr (4%), and Omegle (2%). Adolescents disclosed mental-health issues (18%), that they engage in self-harming behaviors (5%), and they thought about or attempted suicide (4%). They also mentioned that they have other mental health issues such as anxiety, personality disorders, etc. (4%). Adolescents posted on a range of topics, including their offline sexual experiences, relationship advice, hopes and dreams, family, dieting, mental health, self-harm, and suicidal ideation. However, we made the explicit choice to not conduct a person-based analysis that could unintentionally aggregate disaggregated data in a way that would make an individual adolescent identifiable.

Seeking Support for Online Sexual Experiences

Adolescents who sought support spoke generally to the crowd—asking everyone for advice, sharing their intimate personal experiences, or were just venting, so they could be heard. They openly complained about their problems, shared their stories, and recounted awkward situations for which they explicitly sought support and/or advice. We identified four different types of online sex-

ual experiences for which adolescents sought support: 1) Sexting (78% of support seeking posts, N=2,706), 2) Sexual Orientation (16%, N=549), 3) Sexual Abuse (8%, N=292), and 4) Explicit Content (7%, N=237). In the subsections below, we describe each type of support seeking in more depth.

Seeking Support about Sexting

The most prevalent type of sexual interaction for which adolescents sought support was sexting. Of the posts about sexting, 66% involved requests to exchange sexual messages and/or nude photographs or videos. The other 43% discussed cybersex via real-time messaging or video-sharing apps. First, we analyzed these posts to understand the situational context of the experience—whether the adolescent said they were the initiator or recipient of the request. In almost half of the posts (46%) adolescents said that they received a request, rather than being the initiator (19%). Of the sexting posts where the adolescent was the recipient, 61% implied that the exchange was unwanted:

"GAH! I can't believe that I was talking to a guy for 5 minutes and out of nowhere he just sends me a naked pic. THE FUCK?! Honestly, WHY?" 13yr old Female

In cases where users initiated the interaction, they often thought sexting might bring their relationship to the next level. Sometimes, they offered to send nudes to a romantic partner but did not receive the response that they were expecting. Therefore, they sought advice on how to interpret the situation, after it took an unexpected turn:

"I have been going out with my boyfriend for like 5 months now and we haven't done anything past kissing. I offered to send a nude a few nights ago and he got really pissed and said how I shouldn't send nudes to people. we haven't talked since. Confused!?!?" 14yr old Female

In these situations, adolescents (mostly females) were often more concerned about how the sexting interaction changed the nature of their relationships with others, rather than the repercussions from having engaged in sexting itself. Other times, they posted because they sent a nude photo or engaged in cybersex, but then something negative happened (e.g., someone sharing the nude images to others at school) to make them regret their own actions. Youth sought advice on how to recover from these types of mistakes:

"I sent nudes to my friend. I know it's stupid but the compliments were so nice and made me not hate myself for a while, I trusted him... at school he showed half my grade. In so embarrassed I cut when I got home and filled the tub with blood... I hate him so much, but I hate myself more. Please help :(" 15yr old Female

In most cases, when adolescents said they initiated a sexting interaction, they expressed doubt, regret, and confusion. Their posts reflected the need to get feedback from others on how to recover from these situations.

Next, we explored the relationship context between the youth and the people in which the sexual exchange occurred. Most of the posts for sexting were regarding interactions with strangers (37%). Adolescents complained about unsolicited nudes and sexual advances and expressed disappointment that people just wanted to use them as a sexual object, rather than get to know them for who they were. When they talked about sexting requests from strangers, they were more likely to complain that the request seemed out of the blue or random. They were less likely to feel pressured or to reciprocate in the exchange. In contrast, 30% of the posts involved sexting interactions with a romantic interest. Usually, adolescents asked for advice on how to navigate sexting within their relationships. Male adolescents were more likely to ask for advice on why sexting exchanges stopped with a romantic partner, implying they wanted the interactions to continue:

"My girlfriend and I are long distance and before she met me in person we often did sexual things

on Skype. Ever since she went back home we haven't done anything:/ Any advice" 17yr old Male

Meanwhile, many adolescents expressed excitement and nervousness when sexting became a part of their relationships. They were curious about these new sexual experiences and concerned about doing it safely:

"I just sexted with my boyfriend for the first time and he's 14 and I'm 15 but omg. idk. it was actually kinda fun...I haven't sent any nudes. so I'm safe... agh. idk. can't tell anyone cuz then they'd judge me." 16yr old Female

Yet, females were also often frustrated that they were being pressured to share nudes, and scared that if they did not, they would be rejected by their love interest:

"I am worthless. All my bf like me for one stupid thing. Nudes pics. And when I refuse to send them they brake up w me! I want to slit my wrist so bad!" 15yr old Female Females were also confused by their male friends, who made unwanted advances, struggling to set boundaries:

"One of my best friends from school is getting really weird... He was talking about sending me dick pics and now he's telling me he's really hôrny... I've just stopped replying because I don't want to go there with him... He means to much to me... Can I just ignore him?" 17yr old Female We were also interested in understanding how adolescents coped with or responded to these sexting situations. For the posts in which they were recipients, most of them (33%) did not engage in the interaction (i.e., did not send nudes or sexual messages). When the initiator of the request was a stranger (sometimes another user on the platform), it was easier for them to voice their desire not to engage:

"There's a guy I message off here... he keeps trying to sext me and I'm constantly making it 100% clear I don't want to do that" 14yr old Female

Others either blocked or used in-app reporting features to stave off the unwanted request. Often, youth posted their annoyance about having to deal with unwanted solicitations when they were already dealing with other stressors.

"Some annoying fuck boy just sent me a dick pic. I blocked and reported him... this day has been annoying overall and he is the last thing I really needed." 15yr old Female

Some (16%) did not take any actions in response to receiving a sexting request or receiving nudes. This occurred more often when the adolescent knew the person who made the request. For instance, when the initiator was an acquaintance, friend, or romantic interest. In these cases, they often ignored the request or receipt of nude material because they did not want to damage the relationship:

"The guy I really like keeps messaging me, he's really nice but he keeps sending dick pics. I ignore them but he carries on... like wtf do I do?!' 17yr old Female

Meanwhile, 15% of the posts indicated that adolescents actively participated in the sexting exchange after someone else initiated it. Adolescents (mostly females) expressed regret (similar to when they initiated the exchange) and reflected on how they felt pressured. In some cases, they said that others threatened to self-harm or kill themselves if they did not comply with the request to send nudes:

"My friend told me she would kill herself if I didn't send her a nude! I feel used yet again I did it but... Idk what to feel anymore" 14yr old Female

Adolescents also received nude pictures from someone else (usually someone they knew), which made them aroused, so they reciprocated. Then, they sought advice on how to interpret how this might affect their relationships:

*"So I've liked this guy for 6 years and he's like my best friend and we love each other so much but like as friends. He's 14. And he just asked me to send pictures to him. Like naked.... And he sent me a pic of his d**k and it kinda turned me on. I kinda sent him a pic back. What does that mean?"* 14yr old Female

Many of the posts highlighted the complicated relationships within the adolescents' lives, where best friends sometimes tried to cross the line to make the friendship sexual. Often, when they posted about sending/receiving nudes to others, even their romantic partners, they expressed reluctance, ambivalence, and guilt, asking advice as to whether they should be sending explicit photos at all. Peer-pressure emerged as a common theme, where female adolescents felt like they had to exchange sexual images with their love interest if they wanted the relationship to continue:

"He finally said I love you. But then right after he started asking for nudes. And i realized that he only said it so I'd send him nudes. I don't want to but I'm scared he'll break up with me" 14yr old Female

In summary, adolescents often sought support about unwanted sexting solicitations from strangers but struggled the most when these solicitations came from people they knew. Next, we discuss the characteristics of the adolescent posts which sought support for their sexual orientation.

Seeking Support for Ones' Sexual Orientation or Identity

Within the support seeking posts, 16% (N=549) were posts where adolescents were asking advice as they explored their sexual identity. These posts were often made by Lesbian, Gay, Bi-sexual, Transgendered, and/or Queer (LGBTQ) adolescents, or those who were still figuring out their sexual orientation. The situational context of these posts varied significantly from the posts about sexting; in most cases, adolescents were the were initiator (44%) rather than the recipient (21%).

Unlike support seeking posts about sexting, posts about sexual identity and orientation often did not involve another party, so adolescents were simply posting about their own experiences, so that others could help them disentangle their thoughts and emotions:

"So I feel very confused about my sexuality. I've had crushes on guys (I'm a girl) and I've never had a crush on a girl, I think lots of girls are hot/sexy/beautiful ect. And if I see a dick on tumblr, it just freaks me out and I just don't like it at all really. If anyone could help me "find myself" that would be great." 13yr old Female

Adolescents who posted about being the recipient of interaction about their sexual orientation often described a cyberbullying situation (see Consequences for more details). When we coded based on relationship context, 27% of such messages came from strangers. In some instances, random people tried to convince adolescents that they were wrong about their own sexual identity:

"A guy I didn't know messaged me on kik asking me out after talking to me for twenty minutes. When I told him no and that I don't like guys he said that I don't know that and that being gay could be cured anyways. He kept trying to convince me that I wasn't gay. He then proceeded to tell me that I'm just a girl with tomboyish tendencies and that I'm just childish when I told him I was gender fluid..." 17yr old Non-binary Adolescent

In about 26% of the posts, adolescents asked for advice about mixed messages they received from friends and/or acquaintances that made them question their friend's sexual orientation, and sometimes, even their own:

"My best friend sends me nudes and flirts with me a lot when i ask her if she likes girls she says no .but that doesn't make any sense? and i kinda like her.btw im bi wtf help" 15yr old Female

Fewer posts described a situation with a romantic partner (9%). In these cases, adolescents in non-heteronormative relationships sought advice on how to navigate disrespectful questions about

their relationships:

"So Im dating this girl and we're constantly asked 'who's the man in the relationship?' I'm pissed cause 1st off were both females... Neither of us have figured out who is either submissive or dominate in this relationship... I didn't want shit like this." 15yr old Female

In some cases (7%), posters described their challenges regarding their families. They expressed gratitude that they could come to the internet for advice when their families were not understanding. They wanted help on how to come out to their families. These adolescents expressed relief that they were receiving support for their choices and wanted to share that relief with others in the support community. In summary, they used the online peer support platform to explore their sexual identities and get advice on how to interpret their own internal struggles around gender and sexuality, as well as how to deal with how this identity affected their interactions with others, both on and offline.

Support Seeking for Sexual Abuse

A relatively small number of posts sought support for online sexual abuse (8% N=292). Some of these posts were about online sexual interactions leading to rape, while others described cyber-violations that emotionally harmed the adolescent given past sexual abuse. Since the posts were categorized as abuse, adolescents were in all cases the recipient of the sexual interaction, rather than the initiator. Many of these posts were significantly longer than posts in the previous categories, often using a narrative approach to tell their stories, as there was a lot for these adolescents to unpack. In several cases, posts that would have otherwise been classified as unwanted sexting interactions were coded as abuse due to the adolescents' personal history and emotional trauma. Often, someone unintentionally brought back memories of past sexual abuse that made the adolescents feel vulnerable, violated, and triggered:

"EEEEW!!!! some guy just randomly messaged me on Kik and asked if i had a big... i cant even stand to say it... i was raped when i was five and the dude that did that to me... kept saying "look at you, you have a peanut!!!!" #ImScared" 16yr old Male

In other cases, the unwanted interaction was with an adult that the adolescent knew and realized that the behavior was inappropriate. Below, the 16-year-old female asked the community whether she should entrap her predator to get him "sent down," or whether doing this would be wrong:

"There's a man who's is my dads friend... he just asked me to send him a pic of me in my bra. Am I doing a bad thing or carry on and get him sent down? Obviously I have been abused as a child and I hate pedophiles I'm just helping clear the world if a sicko I'm 15 he's 40 he has been messaging me for over a year." 16yr old Female

In terms of relationship context, 45% of the posts involved a stranger. These adolescents recounted situations where they were forced to engage with someone else sexually online, often years after the abuse occurred:

"i was 11 when some 20 something year old man forced me to send him nudes and forced me to do gross things on facetime, telling me he'd kill me if i didn't, or that he'd send a video of me watching him jack off to all my contacts." 16yr old Male

About 21% of the posts involved acquaintances. In these cases, the abuse occurred both online and offline. Like the previous example, posters often felt forced into a situation they could not control. Sometimes this involved past consensual interactions that the adolescent later regretted. They posted because they felt trapped and scared.

"Ive been sexually abused/harassed online and in real life by these guys at my school but i can't actually get help because i... Did some stuff with one of them so if i try to tell someone, they have blackmail on me, and i could potentially get arrested and im terrified." 15yr old Unspecified

Gender

Another 17% of the online sexual abuse posts involved a significant other. Again, these posts often co-mingled offline and online abuse that happened in the past with an "ex." Adolescents felt triggered and needed somewhere to share and these traumatic experiences. A small percentage (7%) of adolescents posts mentioned online sexual abuse involving a family member (e.g., parents, siblings, uncles).

"so last night I asked my brother if he wanted to smoke weed with me he said no so then after awhile he asked for my snapchat I gave it to him he asked me if I wanted to see his dick I said WHAT" 14yr old Female

A concerning trend across the sexual abuse posts was that the poster did not indicate reporting the abuse to the proper authorities. This was often because they felt trapped by their abusers or ashamed by their own actions.

Seeking Support for Sexually Explicit Content

The least common (7%, N=231) type of sexual interaction for which they sought support was sexually explicit content. In about 29% of the posts, they were the recipients of explicit content, while in 23% of the posts they consumed pornography at their own volition. In the posts where adolescents were the recipients, 19% indicated that the interaction was unwanted. In most cases, they complained about unwanted explicit content showing up in their social media feeds from the people they were connected to online:

"I'm sick off it seening half naked, nude, bra and undies on my feed Instagram, Facebook and everywhere. It's so disgusting" 17yr old Female

Adolescents also described situations that occurred with an acquaintance (15%) or romantic partner (15%). In these cases, they were often being pressured to consume sexually explicit content for the benefit of the relationship:

"I have a fear of sex and nudity. My boyfriend is nice he wanted to help me... He asked me to watch porn because he thought watching it would make me less scared. But turns out I got even more scared. " 15yr old Female

In summary, we discussed support seeking for different sexual categories and their characteristics. Next, we discuss the consequences of online sexual experiences.

The Consequences of Online Sexual Experiences

Next, we explored the consequences identified in the posts after adolescents were involved in online sexual interaction.

Negatively Impacted Mental Health

Many posts explicitly or implicitly described how online sexual experience negatively impacted the adolescent's mental health. This was most prevalent within sexual abuse posts (48%), where adolescents described emotional and mental trauma, combined with guilt, shame, and fear:

"I've been raped by a guy I met online. I thought it would never happen to be because I wasn't really insecure... I've been chatting with him for 1 1/2 years before he wanted to meet... It's all my fault. I feel awful." 15yr old Female

This sentiment was also present in 35% of the sexting posts when shared images were later used against them:

"i sent a nude to my now ex bf bout a year ago and he sent it to ppl and every1 keeps calling me a thot and just the worst things u could imagine at school i already planned everything out for suicidie" 13yr old Female

For explicit content, the most common mental health problem (29%) was addictive behaviors that made adolescents feel out-of-control and ashamed:

"Sending nudes is a bad thing? As long as watching porn? I'm a 14 year old girl and I want to stop doing the two of them but I always fail and feel really bad" 15yr old Female

When adolescents revealed their sexual orientation, they were often ostracized or ridiculed by others, which negatively impacted their mental health (24%). Losing the support of friends and family during a critical time in their sexual development was detrimental to their self-esteem:

"Bisexual and ashamed my friend blocked me off snapchat told all her friends they started removing me off their friend list just because I came out I cAnt even look in the mirror of myself" 16yr old Female

Mental health consequences ranged greatly—from mild embarrassment, rage, hopelessness, self-harm, to suicidal ideation and actual attempts on the adolescent's life.

Online Harassment Resulting from Sexual Interactions

Being bullied was also a direct consequence of online sexual experiences. In 15% of the online sexual abuse posts and 4% of sexting posts, insult was added to injury when the adolescents' sexual abuse was recorded and shared with others, who then ridiculed them for being promiscuous or damaged goods:

"I was sexually assaulted my an ex I had and he put my assault on video and showed a bunch of people...which led me to being sexually harassed at school ... the guy is now my ex and he's texting

me saying that I'm a santa and a hoe for being sexually assaulted." 15yr old Female

In some cases, this hurt the adolescents' reputation and led to name-calling. Some more extreme cases often occurred once an intimate or trust relationship degraded to the point where the offending party used the digital imagery as a form of revenge porn or sextortion.

Other posts mentioned being bullied because of sexual orientation (10%). Adolescents complained recounted receiving hate messages from the people who harassed them over their sexual identity or orientation:

"Ok so I'm gay... And some how some random person found that out and messaged me that I'm a stupid ass lesbian and I should go die in a hole... And I'm fuck sick and tired of you people who can't accept other people that are apart of the LGBT+ community." 12yr old Female

Harassment compounded the problems adolescents experienced when trying to understand their sexual identity.

Positive Outcomes of Online Sexual Interactions

Although most posts indicated negative consequences from adolescents' online sexual experiences (at least the ones they posted about in which to garner support), some adolescents expressed positive outcomes. For instance, sexting gave them pleasure or made them feel good about pleasing others when other life-situations made life seem depressing. To this extent, sexting was used as a coping mechanism to get some relief and a sense of empowerment.

"Recently I've found myself wanting to start to send nudes again. I know I shouldn't, it's illegal and over all just a thing that could come back and haunt me. I just feel so out of control with my life and so worthless at the moment that at least if I can give someone some sort of pleasure it'll make me stop feeling this way." 16yr old Female

In the next section, we describe the posts where adolescents were trying to connect with others.

Trying to Connect with Others

In addition to seeking support, some adolescents posted because they wanted to connect directly with others (15% N=635). These posts differed from support seeking posts because the adolescents were explicitly trying to make one-to-one connections for support, rather than seeking it broadly from the crowd. Most of these posts (62%, N=392) were from LGBTQ youth trying to connect with other LGBTQ, while other posts were adolescents trying to make friends (37%, N=234) or to talk to someone else who was also sexually abused (5%, N=30). A common type of post in this category was an adolescent stating their sexual orientation and providing their Kik user ID to have other youths of that orientation contact them. In some cases, adolescents wanted to talk to someone who could help them come out to their families:

"Is anyone else bisexual on here? Need some advice, I don't know how to come out to my family. Could you leave your Kiks below. Thank you" 15yr old Female

Several posts came from frustrated adolescents, who wanted to connect with others on a meaningful level but wanted to avoid sexual solicitations. These posts indicated that even though this platform was for support, adolescents often received unwanted sexual solicitations instead:

"Y'know what I really hate..? When I say on here "I really want to talk to someone." and when I do all they want to do is sext or send dirty pictures. I really do want to just TALK.." 15yr old Female

Females were particularly forlorn that they were not receiving the support they came there to get:

"I just feel so lonely. Nobody really talks to me and when they do talk to me all they want from me is sex… I don't want to talk about sex and sending nudes…I want to talk about my problems without getting shut down about how I'm just complaining just to complain." 16yr old Female

Some male users even acknowledged that most guys seemed to come to the platform to engage in sexual interactions. As a result, "guys" who joined the platform to make friends and talk had a hard time connecting with overly cautious "girls." In the next section, we discuss the posts that aim to provide advice to others.

Giving Advice to Others

There was a small portion of posts (5%, N=200) that gave advice to others about online sexual risks. Of these posts, (62%) gave others advice on how to deal with unwanted sexting solicitations and even bad actors on the platform itself. Posts often included a general plea, an expression of frustration, and/or actionable advice on how to avoid such unwanted interactions. Other posts specifically warned others about users that they had negative interactions with in the past, so that they could stay away from predators:

"Stay the hell away from [ID] he is a perv all he wants is for girls to send him nudes he found my Kik on here and started texting me..." 14yr old female

Other adolescents posted pleas to potential offenders telling them to go elsewhere and not prey on vulnerable people.

Discussion

We discuss the implications of our findings and present design recommendations for online peer support platforms.

Online Sexual Experiences as the New Norm

Exploring one's sexuality is a normal and healthy part of adolescence [104], and as adolescents spend more time online, online sexual experiences have become a normal part of their lives. Adolescents in our dataset often treated online sexual interactions as if they were a natural progression of romantic relationships. Similar to offline sexual interactions [263], adolescents in our study wanted help navigating these situations and were desperately reaching out to strangers for advice on how best to handle serious life decisions. Unfortunately, while there are extensive resources for adolescents to learn about healthy versus risky sexual experiences offline [154], the risk narrative around online sexual interactions has limited sexual health resources for these interactions from becoming a mainstream part of sex education. We recommend a societal shift that acknowledges that online sexual interactions are now part of a adolescents' everyday life; therefore, we need to teach adolescents how to engage in these sexual experiences safely. We urge researchers and practitioners to advocate for safe online sexual education for adolescents. For instance, educators could help adolescents weigh the benefits and risks of sexting and discuss safety and exit strategies.

The Double-Edge Sword of Online Peer Support

Our analysis uncovered that adolescents, specifically, ones struggling with their mental health, sexuality, and technology-mediated relationships, came to this online platform to seek support. Vulnerable youth (e.g., LGTBQ and survivors of sexual abuse) often reached out to others directly for help and to form personal connections based on shared life experiences. In this way, the platform empowered adolescents, gave them hope, a sense of community, and belonging—all positive characteristics that have been associated with online mental health peer support platforms [205]. Yet, while these adolescents were exposing their most intimate thoughts, desires, and personal details (e.g., Kik IDs) to receive support from their network, they also received unwanted sexual

advances from strangers to the point that some adolescents posted to warn others about malicious users. Essentially, the platform was exposing adolescents to some of the same harmful experiences they were there to overcome. This raises the question as to whether the benefits of online peer support outweigh the potential risks—or how we might design these platforms in a way that optimizes benefits and mitigates risks. Therefore, we offer some recommendations for designing online peer support platforms to better meet the needs of adolescents.

Implications for Design

While the anonymity of the platform gave adolescents the opportunity to discuss topics they normally did not feel comfortable talking about in the real world, it also placed them at higher risk. One of the main themes that emerged in this paper was the problem of adolescents encountering "bad actors" [44] as they sought peer support and social connection with others online.A possible solution for combating bad actors is to create safe spaces that are designed specifically for adolescents to have frank conversations about their online sexual experiences. These peer support platforms could be membership restricted by age and real identity but facilitate support through anonymity. Further, online peer support platforms could implement peer-based affinity spaces [13] that are restricted to adolescents who identify with certain groups (e.g., LGBTQ, females, by age, etc.). These affinity spaces might safeguard youth from encountering unwanted solicitations or online harassment about their sexual orientation. Ideally, these platforms would be moderated by trained peer counselors and sexual health professionals that understand the types of sexual interactions teens encounter online, so that parents can be assured that their adolescents are getting sound advice. Making such online resources accessible for adolescents could prevent them from going to less reputable websites or niche websites (such as ones geared towards mental health issues) to seek peer support. We might also reflect on whether designing platforms specifically for safe consensual sexting for adolescents would be beneficial. While the platform studied was moderated and

had clear community standards against online harassment and solicitations, it was clear from the posts that unwanted interactions that violated these standards still occurred. This problem could potentially be alleviated by leveraging machine learning approaches to identify such violations. For instance, Wohn et al. [301, 4] recently proposed leveraging human-technology partnerships and algorithmic systems to alleviate the burden of human moderators. Combining this approach with Wisniewski's and De Choudhury's [3] concurrent efforts to create human-centered algorithms [265] to accurately detect adolescent online risks, could be a novel approach to ensuring the online safety of adolescents seeking online peer support. As such, an interesting design implication and future research problem is determining how human-centered approaches to machine learning could be used to disentangle online sexual abuse from other forms of normal online sexual exploration to improve risk detection. For instance, we found that relationship context was a salient feature in unequivocally unwanted sexual solicitations. Therefore, accurately classifying or allowing adolescents to categorize individuals as known or unknown contacts would help automated approaches distinguish between unwanted solicitations and more nuanced interactions. Such algorithms could then be embedded in peer support and/or social media platforms, so that these platforms share the social responsibility of protecting adolescents from sexual exploitation, while not inhibiting them from seeking support for or engaging in developmentally appropriate sexual behaviors online. Additionally, posts often disclosed sexual abuse, self-harming behaviors, and suicidal thoughts. Similar human-centered risk detection approaches could help identify these imminent risks to youth and offer real-time support and provide evidence-based interventions to mitigate harm.

Limitations and Future Research

Our analysis was based on posts made by adolescents on a mental health-oriented peer support platform for adolescents and young adults. Therefore, our results may not be generalizable to other populations. We likely encountered more negatively biased sexual experiences and abuse

narratives given the nature of the platform. Future research should verify whether our results hold for more diverse adolescent populations. Second, the challenge of scaling big data for qualitative research is an open problem that multiple researchers in the SIGCHI community are currently trying to tackle [95, 5]. Our dataset was scoped based on a grounded and iterative process of manually identifying relevant keywords in the data. It is possible that other relevant keywords exist but were not included in our query. For instance, the word "fuck" was intentionally removed because this sexualized language has become prevalent in youths' everyday discourse and resulted in a high number of irrelevant posts.

Most importantly, how to ethically conduct online research, particularly with vulnerable population [286] is an important open issue within the SIGCHI community that warrants continuous scrutiny [115]. Hallinan et al. [115] explains that there is no single solution to ensuring ethical research and one must think about the beneficence (i.e., benefits versus risks) of one's research "holistically." For instance, we considered the public's expectations of the platform based on the site's terms of service, which stated that the data may be used for research purposes. Yet, we felt this was insufficient protection by itself, which is why we made sure that our quotations were anonymized and not publicly searchable as to mitigate potential harm to the youth whose data we analyzed. For our future work, we plan to take the same care in examining the types of peer support and advice adolescents receive from strangers (i.e., comments on these posts) about these experiences.

Conclusion

The key take-away from this research is that online sexual interactions need to become part of the everyday discourse when educating adolescents about safe sex. We investigated the online sexual experiences of adolescents using real-world posts from a peer support platform. We found that online sexual experiences have become an irrevocable part of adolescents' sexual development

and identified some of the benefits and challenges adolescents encounter when seeking support for these experiences.

CHAPTER 4: STUDY 2: INSTAGRAM DATA DONATION: A CASE STUDY ON COLLECTING ECOLOGICALLY VALID SOCIAL MEDIA DATA FOR THE PURPOSE OF ADOLESCENT ONLINE RISK DETECTION

Citation: A. Razi, A., AlSoubai, A., Kim, S., Naher, N., Ali, S., Stringhini, G., De Choudhury, M., and Wisniewski, P. , "Instagram Data Donation: A Case Study on Collecting Ecologically Valid Social Media Data for the Purpose of Adolescent Online Risk Detection," ACM CHI Conference on Human Factors in Computing Systems, (CHI 2022).

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.

Introduction

In recent years there has been considerable interest e.g., [120, 144, 262] in detecting and/or mitigating [297, 9, 10] 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) [39]. 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 [233, 149, 266]. 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 [233]. 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) [149]. Incorporating the perspectives of victims is crucial, as it enables the machine learning models to catch implicit inferences to the said risk [148]. 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 [233, 149], 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 [90].

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 [116, 40, 60]. 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 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 [300] 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.

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) [23]. Instagram and YouTube are the top social media platforms being used by half of U.S. teens ages 13 to 17 [23]. Instagram provides a way for users to download their data, as

Figure 4.1: Instagram Data Donation Main Page

General Data Protection Regulation (GDPR) [106] mandates social media companies to provide options for users to download their personal data.

Figure 4.1 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, and 5) are willing to share their Instagram data with us for the purpose of research.

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. The consent and assent forms for this study can be found in the appendix as a reference.

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. All the survey questions are displayed in Appendix C.

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. [87]. 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. Social capital is defined as “the sum of the resources, actual or virtual, that accrue to an individual or a group by virtue of possessing a durable network of more or less institutionalized relationships of mutual acquaintance and recognition” (p. 14) [50]. Bridging social capital refers to "week ties" in online connections while bonding refers to "tightly-knit" such as emotionally close relationships [87]. We also utilized **Social Media Disorder Scale** by van den Eijinden [275] which is a psychometrically sound instrument to measure social media addiction.

Negative Online Experiences

Next, we asked questions regarding potentially negative experiences that participants had on Instagram. **Cyberaggression and Cybervictimization (CAV) Scale** [252] by Shapka and Maghsoudi was used to measure cyberbullying experience both as a victim and a perpetrator. We also used questions regarding **Deception of Cyberbullying Victimization and Perpetuation scale** by Doane [79] to understand if participants experienced deception and lying on Instagram. **Youth Internet Safety Survey (YISS) Unwanted Online Experiences** by Mitchel et al. [195] was used to measure sexual solicitation, unwanted exposure to sexual material, and produced sexual images.

Personal Experiences and Demographic

We utilized **The Short Warwick-Edinburgh Mental Well-being Scale (SWEMWBS)** by Tennant et al. [269] to measure well-being and mental-health, **UCLA Loneliness Scale** by Hays et al. [121] to measure loneliness, **Patient Health Questionnaire (PHQ-9)** [161] to measure depression, **Inventory of Statements About Self-injury (ISAS)** by Klonsky and Glenn [152] to comprehensively assess the functions of non-suicidal self-injury (NSSI), and **Risky Behavior Questionnaire for Adolescents (RBQ-A)** by Auerbach and Gardiner [28] to assess risky behavior engagement,

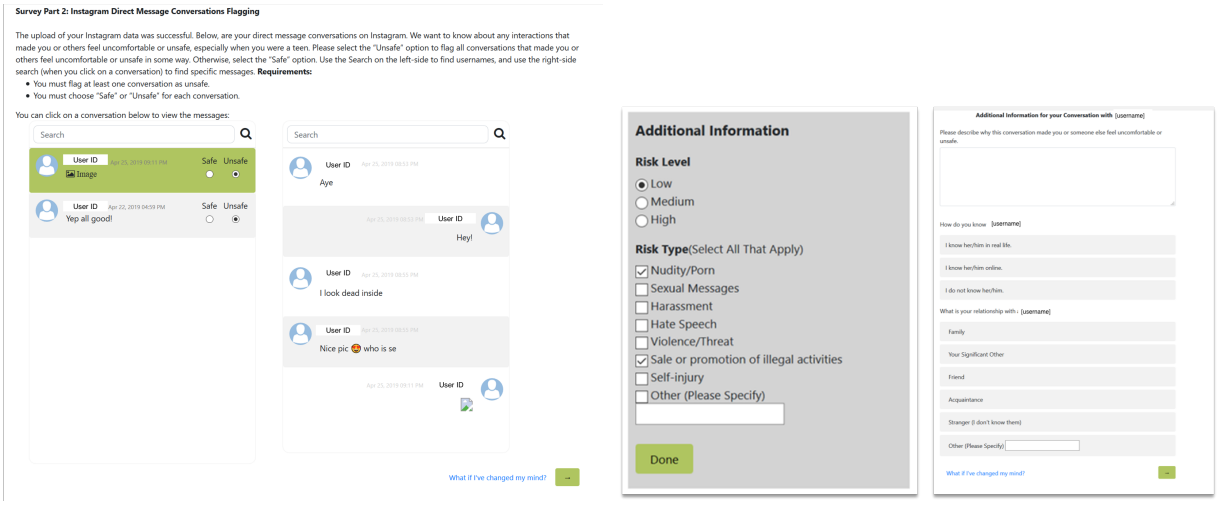

Figure 4.2: Screenshot of (a) Participant Conversation Selection Screenshot (b) Participant Messages Risk-flagging Screenshot.

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. All the measures for the survey were reviewed by Nicholas J. Westers, Psy.D., ABPP, and Board Certified in Clinical Child and Adolescent Psychology.

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 conversations in a sequential fashion, so they could review their interactions and flag each conversation as 'safe' or 'unsafe', displayed in Figure 4.2(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 selecting at least one unsafe message for risk type and risk level as shown in Figure 4.2(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 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.

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

We then grounded risk levels based in the existing adolescent online risk literature [298] 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 were 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 4.2(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 [231].

System Technical Details

Figure 4.3 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 backend 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 every time a new folder was uploaded to the AWS S3 bucket.

- **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.

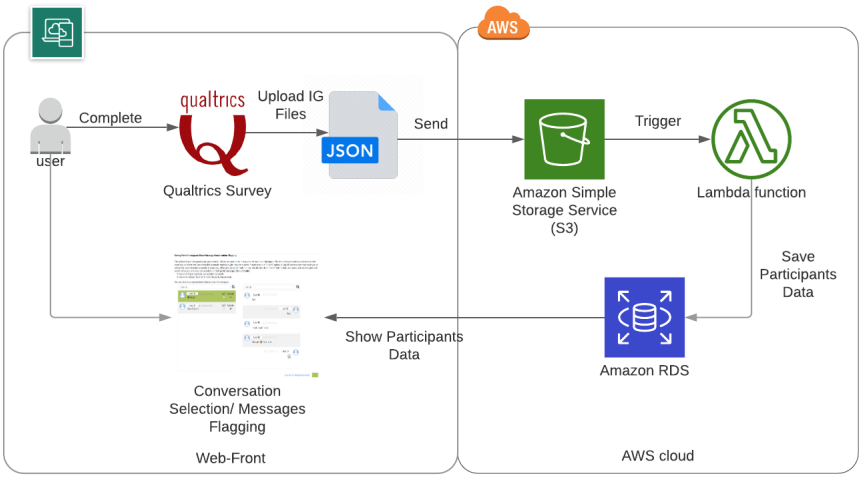

Figure 4.3: Instagram Data Donation System Architecture.

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 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)

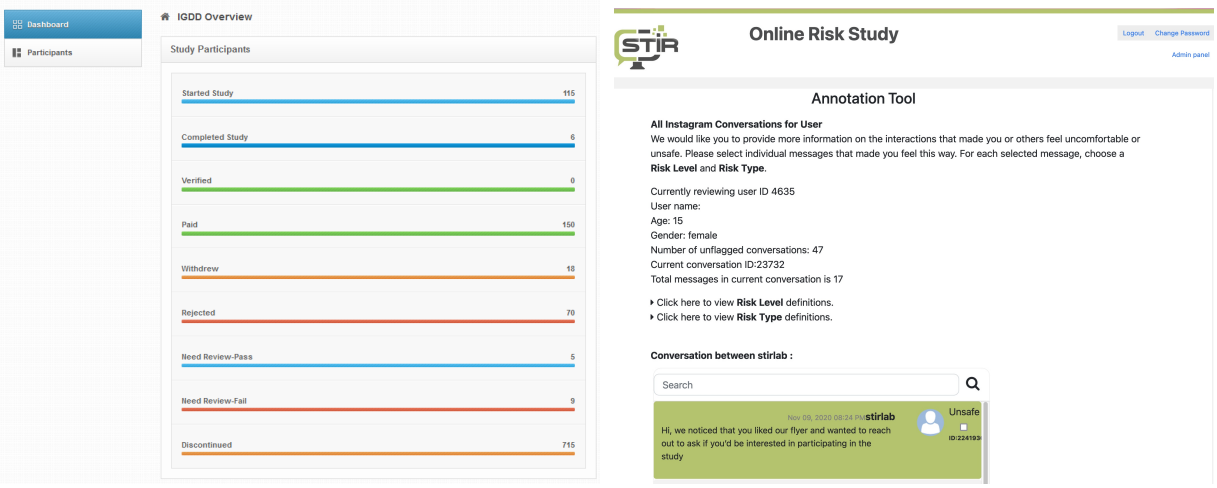

Figure 4.4: Screenshot of (a) Participants Dashboard (b) Annotation Tool.

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.

- 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 the latest available versions.

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 4.4(b)) to facilitate this annotation process. The third party annotators were provided a similar interface to the participants' interface to flag any unsafe

messages.

Data Verification Process

To keep track of the number of participants who participated in the study and ease the process of the data verification we developed a web-tool displayed in Figure 4.1. At the time of writing this paper, 115 youth had started the study and were in route to completing the study. In addition, 150 participants completed the study and passed the data verification quality check. We adopted various quality checks 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 having 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.

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. Here are some descriptive statistics of our collected data: From 150 verified participants 69% 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 = 16, Standard Dev. = 6.2. Most of the participants recognized themselves as heterosexual or straight 47%, however, our dataset also includes 29% bisexual, 11% homosexual, and 13% who preferred to self identify. Next, we found that 41% of our participants are White, 20% Black/African-American, 14% Asian or Pacific Islander, and 6% Hispanic/Latino and 19% belonging to mixed races or who preferred not to self-identify. From the 150 verified participants, we collected 26,734 conversations (average=178 and range=1038-17 conversations per participant), where 2,037 conversations were labeled as unsafe by the respective participants. The total number of messages included in these conversations is over 5.8 million, out of which 2,551 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.

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.

Overcoming Technical Challenges

We developed our data collection system as following the Cambridge Analytica data breach [137], 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.

Leveraging AWS Services

One of the challenges was that Instagram changed the users' folder organization and JSON format multiple times after launching the study [59]. 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.

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. .

Overcoming Privacy and Ethical Challenges

Collecting social media data is a sensitive subject itself [94], 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.

Legal and Ethical Challenges

We disclosed ourselves as mandated child abuse reporters [29] 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.

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 timeline on how long we would keep the data, we made sure to follow the university's, state, and NSF data retention policies.

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://suicideprevention*

lifeline.org), and Child Abuse Hotline (*https://www.childhelp.org/hotline/*).

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 [229, 230, 232] 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 built 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 are 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 the 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.

CHAPTER 5: STUDY 4: SLIDING INTO MY DMS: DETECTING UNCOMFORTABLE OR UNSAFE SEXUAL RISK EXPERIENCES WITHIN INSTAGRAM DIRECT MESSAGES GROUNDED IN THE PERSPECTIVE OF YOUTH

Abstract

We collected Instagram data from 150 adolescents (ages 13-21) that included 15,547 private message conversations of which 326 conversations were flagged as sexually risky by participants. Based on this data, we leveraged a human-centered machine learning approach to create sexual risk detection classifiers for youth social media conversations. Our Convolutional Neural Network (CNN) and Random Forest models outperformed Linear SVM and logistic Regression models in identifying sexual risks at the conversation-level (AUC=0.88). Our experiments showed that classifiers trained on entire conversations performed better than message-level classifiers (AUC=0.85). We also trained classifiers to detect the severity risk level (i.e., safe, low, medium-high) of a given message with CNN outperforming other models (AUC=0.88). A feature analysis yielded deeper insights into patterns found within sexually safe versus unsafe conversations. We found that contextual features (e.g., age, gender, and relationship type) and Linguistic Inquiry and Word Count (LIWC) contributed the most for accurately detecting sexual conversations that made youth feel uncomfortable or unsafe. For example, safe conversations included more words from the LIWC "family" (e.g., sister) category, while sexually unsafe conversations included more from the "friends" (e.g., friend, neighbor) category. Our analysis provides insights into the important factors and contextual features that enhance automated detection of sexual risks within youths' private conversations. As such, we make valuable contributions to the computational risk detection and

adolescent online safety literature through our human-centered approach of collecting and ground truth coding private social media conversations of youth for the purpose of risk classification.

Introduction

In 2020, more than 21.7 million reports of suspected child sexual exploitation were made to the National Center for Missing and Exploited Children's CyberTipline, which increased by 97% compared to the year prior [6]. With the rise in computer-mediated sexual risks, the Human-Computer Interaction (HCI) and Artificial Intelligence (AI) research communities have collectively worked towards understanding how these sexual risks unfold and can be prevented, ranging from in-depth qualitative accounts of sexual victimization [20, 78] to computational approaches for sexual risk detection [262, 171]. For instance, the #MeToo movement [125] gave rise to a body of work where researchers began to detect sexual harassment and/or abuse within public social media posts [108]. The culmination of increased sexual exploitation of youth online and the rise in state-of-the-art computational risk detection approaches for sexual exploitation produce a timely and critical opportunity to leverage the CSCW community's strengths to actively protect youth online.

A recent review of the computational approaches to sexual risk detection synthesized this growing body of literature and called for a more human-centered approach to machine learning (HCML) to move the field forward in a way that would affect real societal impact [233]. For instance, the review highlighted the need for collecting ecologically valid datasets for training robust classifiers to make accurate predictions relevant to real people and contexts. The extant research tended to focus on publicly available datasets, when the most concerning sexual risks such as sexual solicitation and harassment occur in private online spaces like instant messaging and chat rooms [303]. Further, sexual risk classifiers often did not take into account survivors' accounts of their own risk experiences; instead, they often relied heavily on third-party annotators to identify cases of

sexual victimization [233]. As risk is a highly subjective construct [222], quantifiably operationalizing sexual victimization for the purpose of risk detection is difficult without direct input from the individual who experienced it. Finally, existing approaches primarily leveraged linguistic and semantic cues but rarely considered human-centered insights in terms of the contextual factors that have been shown in the literature to increase one's susceptibility to be sexually victimized or groomed [293]. In our case, relevant contextual factors for youth may include developmental (e.g., age), individual (e.g., gender), and relational (e.g., nature of the relationship) factors that have been found to be salient to increase sexual victimization in the adolescent online safety and risk literature [77, 231, 293]. We posit that it is important to take youths' perspectives of their sexual risk experiences into consideration, so that we can identify contextual features important for risk detection. To do this, we analyzed Instagram Direct Messages (DM's) of youth at both the conversation-level (i.e., all messages exchanged in a given private chat) and message-level (i.e., an individual DM) to address the following research questions:

- **RQ1:** *Based on the first-person accounts of youth, what attributes can help us best predict whether sexual risk is present within a private social media conversation?*
- **RQ2:** *a) Can we accurately predict if a given message is sexually risky? b) If so, can we assess its risk severity level?*
- **RQ3:** *a) How are the contextual, linguistic, and semantic features most predictive of sexual risk inform our understanding of the sexual risk behaviors of youth online? b) What are the most common reasons for misclassifications?*

To answer these questions, we collected Instagram data from 150 adolescents between the ages 13-21 and asked them to flag their own private messages for sexual content that made them feel uncomfortable or unsafe. We collected a total of 15,547 private conversations of which 326 conversations were flagged as containing sexual risks by participants. We then created a balanced dataset

of randomly chosen safe versus sexually risky conversations to train a conversation-level sexual risk classifier (RQ1). We tested several machine learning models, and a Convolutional Neural Network (CNN) model outperformed traditional models (accuracy=89%). For traditional models, the Random Forest model that incorporated age, gender, and relationship type as contextual features with linguistic features outperformed other models with an accuracy of 88%. Next, we developed a message-level classifiers for predicting whether a given message contained sexually risky content with an accuracy of 84%, as well as the level of risk posed to the victim (i.e., safe, low, medium-high) with an accuracy of 82% (RQ2).

To answer RQ3, we unpacked how the contextual features and psycholinguistic attributes (based on the Linguistic Inquiry and Word Count, LIWC [220]) played a role in the online sexual experiences of youth. Young adults were significantly more likely to flag conversations as safe, while young teens (between 13-15) and adolescents (16-18) flagged more unsafe sexual conversations. Safe conversations were more likely to be between the participants and family members, friends, or significant others, while unsafe conversations were significantly more likely between participants and strangers or acquaintances. Next, we analyzed the relative importance among LIWC categories and found distinguishing LIWC categories for unsafe and safe conversations. For instance, unsafe conversations contained more words from the "friends" category (e.g., friend, neighbor), compared to safe conversations that contained more words from the "family" category (e.g., sister, daughter). Additionally, an error analysis helped us identify that most of the misclassified instances were due to short conversations that included links or media. Our analysis sheds light on the salient features to leverage in sexual risk detection algorithms, as well as the online sexual risk experiences of youth. Overall, our research makes the following contributions to the Computer-Supported Cooperative Work And Social Computing (CSCW) research community:

- We took great care and effort to create an ecologically valid dataset based on private social

media conversations of youth. The dataset was labeled by youth from their own perspective of sexual risks, spanning incidents and experiences that may have made them feel uncomfortable or unsafe.

- We went beyond identifying sexual predators or detecting sexual harassment in public posts by building classifiers to assess sexual risk in private conversations of youth. In particular, we developed automated machine learning-based detection models that could act as a key element in ensuring online safety of youth and young adults. In addition, we built machine learning approaches to predict the presence and severity of sexual risks in conversations as well as their constituent messages, followed by highlighting their differences.
- Our findings shine a light on the importance of contextual features (e.g., age, gender, and relationship type) in identifying sexually risky conversations, and how automated sexual risk detection models could utilize them for more human-centered risk detection systems.
- We suggest important design implications for computational approaches for detecting sexual risks in private conversations. Additionally, we contribute to the youth online safety by human-insights relevant to youth unsafe sexual interactions.

Related Work

We highlight potential research gaps in the computational sexual risk detection literature that motivate our work and make a case for using human-centered approaches to close these gaps.

Computational Sexual Risk Detection Literature

The majority of computational sexual risk detection research has been conducted in the context of sexual grooming and identification of child sexual predators (75%), sex trafficking (12%), and sexual harassment and/or abuse of adults (12%) [233]. Much of this work started with utilizing traditional ML approaches during the 2012 Sexual Predator Identification competition ran by PAN [1] [134]. After the competition, researchers continued the effort by presenting different traditional models to detect child sexual predators in the PAN-12 data [92]. A relatively smaller subset of the literature adopted deep learning methods for detecting sexual harassment, abuse, or sex trafficking [233]. For instance, researchers compared the performances of deep learning models on a publicly-available dataset "SafeCity," which includes stories for sexual harassment disclosure detection [144, 171]. While several of these studies achieved high performance, most benchmarked their performance based solely on ML metrics (e.g, accuracy, F1-score, recall, precision) [233]. Although these performance metrics are important to evaluate the accuracy of the models, it is important to consider the social interpretation behind the algorithms to thoroughly evaluate the models in real use [35].

Another theme within previous research on sexual risk detection was that most researchers have mainly focused on predicting risk as a binary task (risky vs. non-risky) instead of considering different risk levels [233]. Yet, what we know about risks posed to youth online is that it is a spectrum that can escalate over time [140], rather than a dichotomous state. Thus, some researchers have tried to differentiate risk by differing levels. Ringenberg et al. [238] used Fuzzy Sets for labeling messages for three levels of risks (low, medium, high), and developed Neural Network models that used these fuzzy membership functions of each line in a chat as input to predict the risky interaction. CNN was found in this work as the best model in predicting risk levels. While

[1] A benchmarking activity on uncovering plagiarism, authorship and social software misuse http://pan.webis.de

Seigfried-Spellar et al. [250] classified PJ conversations based on two risk levels for a contact offense which is determined on the model's predicted probabilities of whether the offender showed-up to meet the decoy in the physical world. Therefore, identifying the risk levels could provide in-depth information on the potential degree of the harm to the youth so proper risk mitigation strategies could be used than just a binary classification of whether the risk exists. Therefore, in our study we leveraged machine learning algorithms to be trained on the conversations level to identify the unsafe sexual conversations and went beyond that to train the models on the messages level to identify the risk levels (low, medium, and high) within these messages.

Leveraging HCML to Improve Sexual Risk Detection for Youth

As Artificial Intelligence (AI) has become an irrevocable part of systems that influence peoples' lives, concerns about uncertainty and potential mistakes made by these systems has become heightened [280]. For instance, AI has been used to identify child predators online by developing a deep understanding of the linguistic cues used in the process of sexual grooming [173, 56]. Yet, without an evidence-based understanding of grooming behaviors, risk detection algorithms could be harmful to those classified as alleged predators (e.g., due to false positives) or potential victims (e.g., in the case of false negatives). We address these gaps by taking a Human-Centered Machine Learning (HCML) approach to detect sexual risks encountered online by youth. HCML keeps humans at the center of the design process by taking into account stakeholders' needs, as well as their perspectives. Leveraging practices from HCML [280] is needed to ensure that knowledge about people is used to create robust algorithms that incorporate stakeholders' perspectives and consider potential mistakes/harms that these systems might make [245].

Great strides have been made towards building robust systems for automated detection of sexual risks, but there are several gaps and opportunities for leveraging HCML approaches that we apply

to our research. First, datasets traditionally analyzed for sexual harassment or abuse were mostly based on public posts on social media, such as Twitter [233]. The most popular public datasets used in the literature for identifying sexual groomers, for instance, utilize the Perverted Justice (PJ) dataset[2] and PAN-2012 competition dataset, which was created from PJ with combination of other datasets. The PJ dataset includes logs of online conversations between convicted sex offenders and adult volunteers posing as minors, which is not representative of real-world data from youth. Analyzing public discourse to understand sexual harassment and abuse is another problem, since it is well known that people often behave differently in public spaces than they do privately. Since most sexual risks occur in private channels [303], it is important to examine these interactions.

Second, most sexual risk detection systems have relied on labels that have not been grounded in the victim's perception of risks. Past literature relies heavily on third-party annotations [233], although their perspective of sexual risks might be different than the actual victims. For instance, Kim et al. [148] has found that ML models for detecting cyberbullying instances trained based on the perspectives of the individuals involved in bullying (i.e., "insiders") outperformed the models trained on data annotated by third-party annotators (i.e., "outsiders") by detecting implicit references to bullying. Thus, incorporating first-person perspectives in ground truth labeling of one's sexual risk experiences is an important step towards establishing a risk detection system that does not estrange the key stakeholders of the system.

Finally, the majority of studies on online sexual risk detection have primarily focused on the textual features which represent the linguistic style embedded in the text [233]. The dominance of these textual features such as N-grams, bag-of-words (BoW), and word embeddings entail the *what* and *how* of the dataset; however, this falls short of encompassing the crucial question of *who* is involved in the specific conversation, post, or comment. Since different people perceive differently based

[2]http://www.perverted-justice.com/

on their life experiences [112], it is important to incorporate the human-centered features, such as age and gender of the individuals receiving the message (i.e., our participants), in the training of risk detection systems. Therefore, identifying and utilizing the social and psychological patterns as indicators of risks may be beneficial in developing more effective risk detection models.

In summary, our work takes a HCML approach to address some of the gaps outlined above to advance computational approaches to sexual risk detection for youth. First, we constructed an ecologically valid dataset that is composed of private conversations donated from youth participants. We go beyond detecting sexual predators to detect sexual risks experienced by our participants at the conversation and message levels. Each participant of the study labeled their own data, providing us with the survivors' perspectives of the unsafe or uncomfortable sexual risk experience. Thus, we focus on sexual conversations that made our participants feel uncomfortable or unsafe, leveraging their first-person account of the experience. This is a significantly different goal than attempting to identify sexual predators. Built upon this ecologically valid dataset and labels, this paper also incorporates human-centered features (participant age and gender, and the relationship between the participant and the offender) in developing an automated sexual risk detection system for youth. Next, we took a human-centered approach by qualitatively analyzing instances of our top performing features to shed light on the sexual risk experiences of the youth participants in our dataset. Specifically, we conducted a feature analysis to quantitatively and qualitatively understand how our feature set not only contributed to our prediction accuracy, but also to better understand the experiences of our participants. Finally, we conducted an error analysis to pinpoint areas of weakness that should be addressed in future work.

Dataset

We used the Instagram Data Donation dataset explained in the previous chapter to develop machine learning models for sexual risk detection. For this chapter, we included risk flags specific to our pre-defined category of "Sexual Messages or solicitations" defined for participants as "Sending or receiving sexual messages. Being asked to send a sexual message, revealing/naked photo." in our analysis. Please note that we used the term "risky" for uncomfortable or unsafe conversations throughout the paper. Participants were also asked to provide more contextual details about each unsafe conversation; for instance, whether the other party in the conversation was an acquaintance, a friend, a significant other, or a stranger. Because of our understanding that pre-existing relationships affect responses in online sexual experience incidents, we considered the knowledge of this relationship relevant to these risk situations [231].

Participants Demographics

Our study was comprised of 150 participants between the ages of 13 to 21 (Av.=16 yrs, Std.=6.2). To recruit a diverse subset of participants, we promoted our study on social media and contacted more than 650 youth-serving organizations. Our participants were mostly female (Approx. 69%) with 21% identifying as males, 9% non-binary and the rest of the participants choosing not to provide their gender. The majority of our participants were heterosexual or straight (47%); however, a relatively large percentage of our participants identified as bisexual (29%), homosexual (11%), or preferred to self-identify (13%). The race distribution of our participants was as follows: Caucasian/White (41%), African-American/Black (20%), Asian or Pacific Islander (14%), and Hispanic/Latino (6%), and 19% belonging to mixed races or who preferred not to self-identify. We had a representation from the following states: Florida (15.8%), California (12.5%), Indiana (2.6%), and 28 other U.S. states. Participants reported that they used Instagram several times a

Table 5.1: Proportion of safe ($N = 13,610$) and unsafe ($N = 20,33$) conversations across contextual factors

Contextual Factors	Factor	Safe (#)	Safe (%)	Unsafe (#)	Unsafe (%)
Gender	Female	10314	76%	1517	75%
	Male	2281	17%	386	19%
	Non-Binary	822	7%	116	7%
Age Groups	Ages 13-15	2980	22%	536	26%
	Ages 16-18	7585	56%	1203	59%
	Ages 19-21	2949	22%	294	14%
Relationship Type	Stranger	7551	56%	1431	73%
	Acquaintance	732	5%	341	17%
	Friend	2504	18%	168	9%
	Significant Other	755	6%	12	1%
	Family	2035	15%	18	1%

day (51%), every day or almost every day (22%), several times an hour (19%), once or twice a week (4%), less than once a month (2%), and less than once a week (1%). Table 5.1 listed the distribution of the safe (N=13,610) and unsafe (N=2,033) conversations by participants' gender, age, and the relationship type.

Characteristics of the Instagram Data

We collected a total of 15,547 Instagram DM conversations from 150 participants (average=178 and range=min:17-max:1038 conversations per participant). The total number of messages shared in these conversations was more than 5 million. A total of 2,033 conversations were labeled by participants as unsafe, and 326 out of these unsafe conversations were labeled as making them feel sexually uncomfortable or unsafe. These unsafe sexual conversations included 44,099 messages belonging to 150 participants. Of these messages, participants flagged 504 messages as 'sexual messages/solicitation' with 44.6% categorized by participants as low, 33.3% as medium, and 22.0% as high risk. Most participants said that the sexual unsafe conversation was with someone

they did not recognize 73%, some knew them from real life 14.4%, and some knew them online 10.1%. Table 5.1 showed that most of the unsafe sexual online conversations of youth were with strangers. Out of these 326 conversations, 26 conversations were group conversations that involved the participant and multiple others.

Methods

In this section, we discuss the data pre-processing and machine learning approaches that we used. We adopted both traditional supervised learning approaches and deep learning models to predict unsafe sexual conversations and risk level of unsafe sexual messages.

Data Pre-processing and Preparation

During the data pre-processing phase, punctuation marks, hyperlinks, stop words, non-latin words, single/numeric characters, and conversations that had less than three words were removed. Emojis were converted to their associated word representations through the demoji Python library[3] to conserve the semantic meaning depicted through emojis.

To train our conversation-level classifiers, we first created a dataset that had a 50-50 split between the conversations that were labeled with sexual risk and those that were not labeled with any online risk. In applications such as risk detection that datasets are usually imbalanced toward having more safe samples rather than unsafe samples, it is common to balance the data between classes [64, 273]. We used random under-sampling to reduce the number of safe conversations to gain an equal number of class samples and create a balanced dataset. After the data cleaning, our dataset contained 264 sexual unsafe and 249 safe conversations (Total=513).

[3]demoji - https://pypi.org/project/demoji/

For classifying the risk level of unsafe sexual messages we had 3 classes with uneven samples (Low= 136, Medium= 65, High= 33). We encountered two issues with this data, first uneven samples and second lack of data. Thus, we oversampled the unsafe sexual messages with risk levels to the class with the highest instances (N=136) to create a balanced dataset. We used the RandomOverSampler[4] library which over-samples the minority classes by picking samples at random with replacement. After oversampling, we had a balanced number of samples for each class (Low=136, Medium=136, High=136), but still did not have enough data for classifiers to perform well. Therefore, we used a Contextualized Word Embedding augmentation by NLPaug Python library [5] Contextual Augmentation called BertAug [153] and doubled the number of instances in each class (N=272). Classic word embeddings might not fit some scenarios since they use a static vector to represent the same word with different meanings. Meanwhile, contextualized word embeddings consider surrounding words to generate a vector under a different context to solve this issue [6]. BertAug provided insertion which is predicted by the BERT language model which is better than picking one word randomly [153]. Also substitution uses surrounding words as a feature to predict the target word. We leveraged augmentation because it is useful in several aspects, including minimizing label effort, lowering the usage of real-world data in sensitive domains, balancing unbalanced datasets, and increasing robustness against adversarial attacks [37].

Feature Engineering

We developed five categories of features for our traditional supervised learning models. We used the features below to build sexual risk detection classifiers for detecting unsafe sexual conversations (RQ1). For detecting unsafe sexual messages and their severity (RQ2), we trained the

[4]https://imbalanced-learn.org

[5]https://github.com/makcedward/nlpaug

[6]https://towardsdatascience.com/data-augmentation-library-for-text-9661736b13ff

traditional models with combination of the five feature types. For feature analysis, we trained and tested the sexual risk conversation classifiers with each category separately, as well as using all five together (RQ3). We used the flagged messages of the participants as ground truth to train these models. Each conversation/message was represented as a vector where each element was the value of one feature. In the following list, we describe the features that compose each category in more detail:

- **Contextual Features (Age, Gender, and Relationship):** We acquired age and gender of the participants from our survey questions. We examined them as features (3 options for gender and an integer for age) for our model since many empirical studies emphasized the role of age and gender in online sexual risks [73]. For the unsafe conversations, participants were also asked the nature of the relationship between themselves and others involved. Based on this participant annotated data, we trained a Convolutional Neural Network (CNN) model (Avg. AUC=0.90) based on concurrent work [anonymized for review] to machine label the safe conversations for relationship type (i.e., stranger, acquaintance, friend, family, significant other). The relationship feature was a categorical number representing the relationship type for each conversation.
- **Psycholinguistic Attributes (LIWC):** LIWC is commonly used to obtain psycholinguistic features embedded in text as well as to quantify meaning across various dimensions [220]. Based on prior work [72], we selected 50 categories spanning across *affect, cognition and perception, interpersonal focus, temporal references, lexical density and awareness, biological concerns*, and *social/personal concerns* and used them as features. To calculate each feature, we normalized the word counts related to each category by the length of the conversation.
- **Sentiment:** Emotion was represented as a sentiment score, which were extracted through Stanford CoreNLP's deep learning tool [179]. The tool gave us a single label that indicated

whether the conversation was positive, negative, or neutral. The scale for sentiment values ranges from zero to four. Zero means that the sentence is very negative while four means it's extremely positive.

- **Term Frequency-Inverse Document Frequency (TF-IDF):** TF-IDF scales down the term weights of terms with high collection frequency by reducing the weight of a term by a factor that grows with its collection frequency [249]. We defined each conversation as a document, and calculated the 25 words with the highest TF-IDF in all conversations. We then used these 25 words as features, calculating the normalized count for them in each conversation.
- **Sexual Lexicon:** To capture domain-specific signals as features, we used a lexicon developed in prior work [231] including 98 words. For each of the words, we use its normalized count in a conversation as a feature.

Machine Learning Models

We chose Linear Support Vector Machine (SVM), Random Forest (RF), and Logistic Regression (LR) as traditional classification approaches. Next, we implemented an end-to-end Convolutional Neural Network (CNN) from the model by Kim [150]. CNN techniques have shown promising results for text classification in recent years [150]. This allowed us to compare the results of the CNN model with the traditional models. Kim [150] made a small modification to the CNN model architecture by Collobert et al. [69]. This architecture involves a convolutional layer with multiple filter widths and feature maps, then a mac-over-time pooling operation [69], lastly fully connected layer with dropout and softmax output. In order to convert conversations to vectors of tokens as input of the CNN model, we used the Keras Tokenizer Python library [7]. We also experimented using pre-trained GloVe [221] to convert text to word embeddings, which capture

[7]keras - https://keras.io/api/preprocessing/text/

the semantics and syntax of words in text. We then built a CNN model that aims to predict whether a conversation/message is sexually unsafe and risk severity of a message. We used participant flagging or annotation as ground truth to train and evaluate this classifier.

Evaluation

We used the average accuracy of the models, standard deviation of the accuracy, F1-measure, area under the receiver operating characteristic curve (AUC), and class-specific precision and recall to evaluate our models on the test sets. We used grid search and stratified k-fold cross-validation ($k = 10$) to tune the hyper-parameters during the training and validation phases. While the accuracy and F1 scores return the general performance of the models, precision and recall of each class and AUC provide more detailed insights. In our study, in addition to different computational measures, we analyzed misclassified instances to understand the reasons why the model failed on those instances.

Results

First, we present the results of the classifiers that predict whether sexual risks at the conversation-level (RQ1). Next, we provide the results of the message-level classifiers that predict the presence of sexual risks (i.e., binary classifier) and the classifiers that determined the risk severity level (i.e., safe, low, and medium-high) of a given message (RQ2). An analysis on the top features that contributed to the best accuracy performance of the conversation sexual risks classifiers (RQ3) is also presented, followed by an error analysis of our classifiers.

Conversations-level Sexual Risk Detection (RQ1)

We implemented and evaluated multiple classifiers detecting sexual risks at the conversation-level, including Support Vector Machine (SVM), Random Forest (RF), Logistic Regression (LR), and Convolutional Neural Network (CNN) using the 476 conversations as training and test datasets. Minimal pre-processing of data improved performance for the traditional models. For instance, while we experimented with lemmatization and spellchecking, traditional models using these resulted in relatively poorer performance because pre-processing removed contextual and linguistic style information from the original conversations. Further, for the CNN model, recall that we experimented using pre-trained GloVe [221] to convert text to word embeddings. However, the CNN performed better with the Tokenizer, because GloVe had a disadvantage with out-of-vocabulary words from the corpus that it was pre-trained with.

Table 5.2: Model performance across different feature sets for traditional machine learning models.

Linear SVM						
Features	**Classes**	**Prec.**	**Rec.**	**F1**	**AUC**	**Accr.**
Combined + Contextual	Sexual	0.77	1.00	0.87	**0.84**	**0.85**
	Non-sexual	1.00	0.68	0.81		
Combined	Sexual	0.73	0.65	0.69	0.70	0.71
	Non-sexual	0.69	0.77	0.73		
LIWC	Sexual	0.83	0.77	0.80	0.76	0.77
	Non-sexual	0.70	0.76	0.73		
Sentiment	Sexual	0.62	0.53	0.57	0.57	0.57
	Non-sexual	0.53	0.62	0.57		
TF-IDF	Sexual	0.73	0.77	0.75	0.67	0.69
	Non-sexual	0.63	0.57	0.60		
Sexual Lexicons	Sexual	1.00	0.05	0.10	0.53	0.49
	Non-sexual	0.47	1.00	0.64		
Random Forest						
Features	**Classes**	**Prec.**	**Rec.**	**F1**	**AUC**	**Accr.**
Combined + Contextual	Sexual	0.86	0.93	0.89	**0.88**	**0.88**
	Non-sexual	0.91	0.84	0.87		
Combined	Sexual	0.90	0.70	0.79	0.81	0.81
	Non-sexual	0.74	0.92	0.82		
LIWC	Sexual	0.74	0.74	0.74	0.68	0.69
	Non-sexual	0.62	0.62	0.62		
Sentiment	Sexual	0.83	0.61	0.70	0.71	0.69
	Non-sexual	0.59	0.81	0.68		
TF-IDF	Sexual	0.83	0.81	0.82	0.78	0.79
	Non-sexual	0.73	0.76	0.74		
Sexual Lexicons	Sexual	0.75	0.41	0.53	0.63	0.64
	Non-sexual	0.60	0.87	0.71		
Logistic Regression						
Features	**Classes**	**Prec.**	**Rec.**	**F1**	**AUC**	**Accr.**
Combined + Contextual	Sexual	0.76	0.96	0.85	**0.82**	**0.83**
	Non-sexual	0.94	0.68	0.79		
Combined	Sexual	0.78	0.67	0.72	0.73	0.73
	Non-sexual	0.69	0.80	0.74		
LIWC	Sexual	0.77	0.77	0.77	0.72	0.73
	Non-sexual	0.67	0.67	0.67		
Sentiment	Sexual	0.59	0.53	0.56	0.54	0.54
	Non-sexual	0.50	0.56	0.53		
TF-IDF	Sexual	0.79	0.61	0.69	0.68	0.67
	Non-sexual	0.57	0.76	0.65		
Sexual Lexicons	Sexual	0.89	0.42	0.57	0.68	0.66
	Non-sexual	0.58	0.94	0.71		

Table 5.3: Performance of CNN

Features	**Classes**	**Prec.**	**Rec.**	**F1**	**AUC**	**Accr.**
CNN						
Language Tokens	Sexual	0.80	0.92	0.86	**0.88**	**0.89**
	Non-sexual	0.95	0.86	0.90		

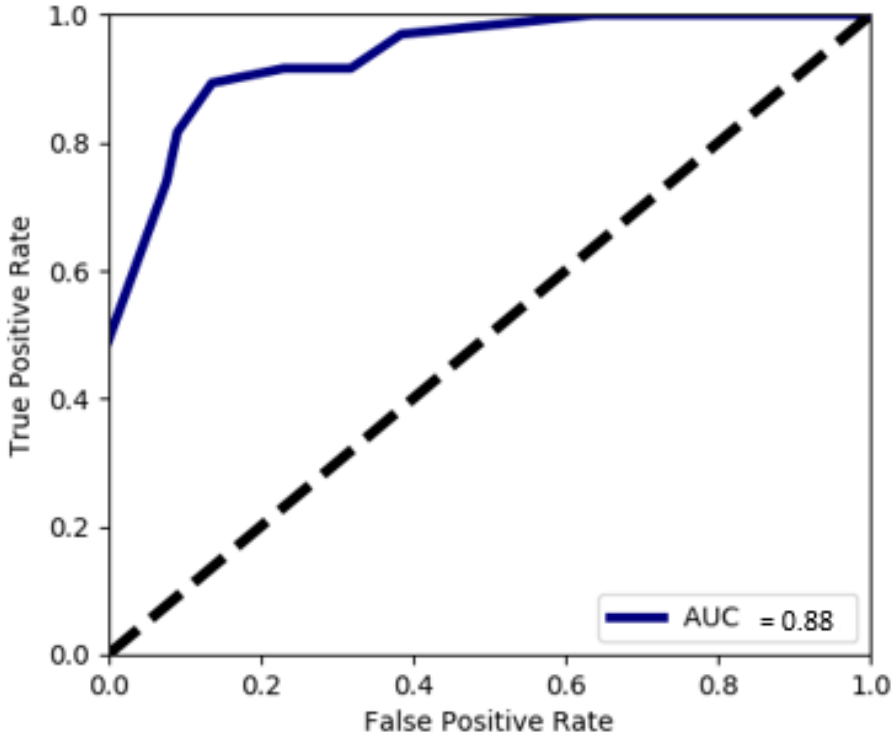

Figure 5.1: CNN Sexual Risks Conversation Classifier ROC

Table 5.2 summarizes the performance metrics of the traditional machine learning models with different feature sets, and Table 5.3 presents the performance of the CNN classifier. Overall, we found that the RF model with combined plus contextual features outperformed other traditional classifiers with an AUC=0.88 and accuracy=0.88, which was similar to the CNN model (ROCs displayed in Figure 5.1). Also, it achieved high class specific precision and recall. For the sexual class, the recall (0.93) was higher than precision (0.86) and for the non-sexual class the precision (0.91) was higher than the recall (0.84).

To analyze the effectiveness of each feature as an indicator for conversation that contained sexual risk, we compared the performance of the traditional models trained separately on each of the

Table 5.4: Sexual Risks Message Classifiers' Performances

Classifiers	Classes	Prec.	Rec.	F1	AUC	Accr.
SVM	Sexual	0.89	0.74	0.81	0.83	0.84
	Non-sexual	0.81	0.92	0.86		
RF	Sexual	0.94	0.69	0.79	0.82	0.84
	Non-sexual	0.79	0.96	0.87		
LR	Sexual	0.89	0.74	0.81	0.83	0.84
	Non-sexual	0.81	0.92	0.86		
CNN	Sexual	0.83	0.85	0.84	0.85	0.82
	Non-sexual	0.80	0.79	0.80		

aforementioned features, as well as all features combined. Overall, we observed that all traditional classifiers had the best performance with the combination of features (LIWC, Sentiment, TF-IDF, and Sexual Lexicons) plus contextual actors (age, gender, and relationship) features, compared to having features separately or the combination of features without the contextual features. This indicated the importance of having contextual features. After RF, Linear SVM resulted in higher performance compared to the LR classifier with (AUC=0.84) and (accuracy=0.85). After combination plus contextual features, the SVM model performed best with LIWC feature (AUC=0.76) and (accuracy=0.77), RF with combined features (AUC=0.81) and (accuracy=0.81), and LR with combined features (AUC=0.73) and (accuracy=0.73).

Message-Level Sexual Risk Detection (RQ2)

In this section, we presented the results of the classifiers for detecting sexual risks in messages and then detecting risk severity of a given message. Given that the traditional models using combined features plus contextual features performed the best for our conversational-level classifiers (RQ1), we also used this approach when building out message-level classifiers (RQ2). Moreover, we tested the message-level classifiers with each feature and combination of features, and combined features

Table 5.5: Message Risk Severity Classifiers' Performances

Classifiers	Classes	Prec.	Rec.	F1	AUC	Accr.
SVM	Safe	0.75	0.90	0.82	0.72	0.62
	Low Risk	0.48	0.25	0.33		
	Medium & High Risk	0.56	0.69	0.62		
RF	Safe	0.81	0.88	0.84	0.74	0. 66
	Low Risk	0.61	0.32	0.42		
	Medium & High Risk	0.56	0.76	0.64		
LR	Safe	0.76	0.88	0.82	0.72	0.63
	Low Risk	0.50	0.23	0.31		
	Medium & High Risk	0.56	0.76	0.64		
CNN	Safe	0.80	0.83	0.82	0.88	0.82
	Low Risk	0.79	0.77	0.78		
	Medium & High Risk	0.88	0.86	0.87		

plus contextual features performed the best, so it was only included for the purpose of parsimony. This enabled us to directly compare the performance of our message-level classifiers to our best performing conversation-level classifiers.

Binary Classification

While our conversation-level classifier performed well, message-level detection was necessary for real-time risk detection and mitigation. Therefore, we trained classifiers for detecting sexual risks at message-level to compare the results with the conversation-level classifiers (displayed in Table 5.4). CNN model outperformed the traditional models with AUC=0.85 and accuracy=0.82. CNN and RF performed better in conversation-level rather than message-level. Though, SVM and LR were slightly better in message-level compared to conversation-level.

Classification by Risk Level

Since participants flagged the risk severity levels (i.e., low, medium, and high) of sexual messages, we trained classifiers to detect the risk severity level of messages in unsafe conversations. Identifying risk levels can be helpful in the process of real-time risk mitigation, which we will reflect on more in the Discussion section. For classifying risk levels, we filtered the original dataset to include the unsafe sexual conversations and trained the risk level classifiers for messages within these conversations flagged by participants for low (N= 136) and combined medium and high (N=98) risk levels (due to the smaller numbers). We randomly selected an equal number of messages from safe conversations (N=234) to classify the messages into safe, low, and medium-high risk levels. We used oversampling (explained in the Method section) to make the classes with lower number of samples equal to the larger sample (each class N=234). We used this approach to make a balanced dataset since participants flagged fewer number of messages as high and medium risks compared to low risk. We trained traditional classifiers with combined features and the CNN with language tokens classifier and compared the results demonstrated in Table 5.5. The results of the unsafe sexual message-level classifiers are shown in Table 5.5. The CNN classifier outperformed SVM, RF, and LR and resulted in AUC=0.88 and accr=0.82 evaluated by 10-fold cross validation.

Contextual Features and LIWC Analyses (RQ3)

Now we unpack how the combination of features plus contextual features and LIWC features performed better in our models and use these results to gain further insights into the sexual risk experiences of youth. We completed this analysis at the conversation-level rather than the message-level since at conversation-level we had more information and more context.

Contextual Features (Age, Gender, and Relationship Type).

Since the combination of features plus contextual features yielded the best performance for the models comparatively, we further analyzed this data to uncover patterns. First, we dug deeper into the contextual features.

For age, a χ^2 test indicated a significant difference between age groups ("Between 13-15", "Between 16-18", "Between 19-21") and their conversation flagging (safe / unsafe) behavior $\chi^2(df = 2, N = 15,547) = 63.33, p < 0.001$. Post hoc testing revealed that younger teens (ages 13-15) were significantly different from older adolescents (ages 16-18) ($p = 0.02$) and young adults (ages 19-21) ($p < 0.001$). There was also a significant difference between older teens (ages 16-18) and young adults (ages 19-21) ($p < 0.001$). The proportions of safe and unsafe conversations as shown in Figure **??** indicated that young adults (ages 19-21) were more likely to flag their conversations as safe, while younger teens (ages 13-15) and adolescents (ages 16-18) were more likely to flag their conversations as unsafe.

Regarding gender, we could not reject the null hypothesis based on the χ^2 test for the gender of adolescents and their risk-flagging behavior $\chi^2(df = 2, N = 15,547) = 5.68, p = 0.058$. This showed that both males, females, and participants who did not specified their gender shared similar conversations flagging patterns.

For relationship type, a χ^2 test showed a significant difference between the relationships types (Stranger, Acquaintance, Friend, Significant Other, Family) and the conversations flagged as safe versus unsafe ($\chi^2(df = 4, N = 15,547) = 882.37, p < 0.001$). According to the post hoc analysis, there was a significant difference between Friend and Significant Other ($p < 0.001$). The post hoc test also found significant differences between Strangers and all other relationship types (Acquaintances ($p < 0.001$), Friends ($p < 0.001$), Significant Others ($p < 0.001$), and Family ($p < 0.001$).

For Acquaintances, we found significant differences with Friends ($p < 0.001$), Significant Others ($p < 0.001$), and Family ($p < 0.001$). Regarding Family, there were significant differences from Friends ($p < 0.001$) and Significant Others ($p = 0.001$). Overall, the proportions of the safe versus unsafe conversations showed that participants were most likely to report having safe conversations with family members, friends, and significant others; meanwhile, unsafe conversations were most likely with strangers or acquaintances, as shown in Figure **??**.

LIWC Categories.

A benefit of linear SVM was that it is an interpretable model to find the most contributing features by looking at the model's coefficients. Therefore, we selected the next best performing feature for SVM, which was LIWC. Therefore, we chose the next best performing feature based on the SVM AUC to further examine the linguistic contributing factors; as the Linear SVM trained on the LIWC had the highest AUC, we looked at the top 15 LIWC psycholinguistic categories in terms of their importance given by the model, as shown in Figure **??**. For instance, the top predictive features for safe conversations contained more first person plurals, such as *"we are,"* which could signal a sense of group identity or togetherness [267]. For example, in a conversation that a 16-year-old female participant had with her friend, they were talking about joining a club, and they used first person plurals frequently to reference their collective action:

> **Other Person:** *Apparently you need clubs to be in honor society*
> **Participant:** *We should go ask what you need to make a club and make it :)*
> **Other Person:** *Haha make our own rules.*

In contrast, the language in unsafe sexual conversations included indefinite pronouns, such as *"It," "it's," and "those"* more often. These linguistic cues tended to show more interest in objects and

things [267]. For example, in a conversation that a 15 year-old female participant described as an *"unwelcome advance:"*

> **Other Person:** *Well I mean I haven't seen you in leggings in a while so idk... Just letting you know I might wanna touch it.*
> **Participant:** *That's really not ok. That's crossing a boundary. I am not ok with that.*

As illustrated above, it was less likely for participants to use collective, first person plural language when they sought to separate themselves from offensive behavior, and offenders used indefinite pronouns to objectify their victims.

For LIWC social processes category, safe conversations included words about family, but unsafe sexual conversation included more linguistic cues about friends. Safe conversations were mostly daily interactions with known others, where it was commonplace to make a reference to a family member. For instance, in a safe conversation between a 16-year-old female and her friend, they talked about regular things that happen with their family:

> **Participant:** *Ah, I'm using my moms charger and it feel so nice to be able to move around without worrying about whether it's charging or not lol*

In unsafe conversations, however, there were more friend-related words, even when the conversation was with a stranger. This was often because the stranger was trying to become familiar with the participants or propositioned them to a *"friend with benefits"* or *"sugar baby"*. In the following unsafe conversation, a stranger sent an unwanted sexual solicitation of this nature. Our 18-year-old female participant did not respond to the unwanted advance. One could see that the text also included indefinite pronouns:

Other Person: *Omg, you are so incredibly beautiful. Hi, my name is X and its an absolute pleasure to meet you. What exactly is it you are looking for? How do you feel about friends with benefits? I'm looking for a special friend to take care of financially as in pay your bills, take you shopping or whatever. You feel you might be interested in something casual like this?*

For affective processes, which was a LIWC category for emotionality [267] *"Happy," "cried," and "abandon"*, unsafe sexual conversations included more negative emotions, such as anger and swear words. For instance, there were lots of profanity, sexual words, and negative emotions in group chats, usually among males. An 18-year-old male participant described one conversation that made him feel uncomfortable or unsafe when he was 16-year-old as:

Participant Description: *They were sexual messages and messages about self harm sent to me at a young age by people i did not know in real life.*

Other Person: *Guys I'm so stressed. History is fucking me in the ass!*
Other Person 1: *Thought that was my job*
Other Person 2: *I wanna kill myself.*
Other Person 3: *I'm sick. Wtf was I doing before I went to bed. Idk probably watching porn or some shit.*

The unsafe sexual conversations also included words from biological processes, such as *"eat," "blood," or "pain"* and sexual categories such as *"horny," "love," or "incest"*. Below is an example from a 21-year-old female that contained sexualized language and referenced body parts:

Other Person: *Hi beautiful My ex baby sent me feet pics, lingerie pics. Texted regularly and FaceTimed in exchange for a weekly $200 allowance and a $800 monthly*

shopping voucher It's more like a companionship type relationship while I'm away. nothing sexual Would you be open to such?

Our participant contextualized the exchange as follows:

> **Participant Description:** *I do not know who this person is. It seems likely to be a bot or phishing scam.*

In addition, sexual conversations included more LIWC perceptual processes such as *"hearing," "feeling," or "seeing,"* which referred to emotional and physical sensations. In contrast, safe conversations included more words from cognitive processes, such as discrepancies (*e.g. "should," "would," or "could"*) and tentative (*e.g. "Maybe," "perhaps," or "guess"*).

Error Analysis.

Next, we looked into specific prediction instances to provide more insights on the factors that contributed to misclassifications. We qualitatively investigated the linguistic style used in the misclassified conversations. First, most of the false negative instances included images or videos, which our text-based models were not able to identify. An example of such instances was a conversation in which someone shared a sexually explicit drawing with a 22-year-old female participant, who described it as *"drawings based on real nude photos of others that I did not consent to seeing"*. The participant in the conversation asked *"all based on real nudes I'm guessing?"* Although this sentence included *"nudes,"* it was paired with the word *"guessing"* from the *tentativeness* category; therefore, the conversation was classified as safe. Other false negative samples were often instances where the participant was added to a group chat with a sexual title and sexual links shared between group members. Since these conversations were short and only included sexual

links/media and a sentence naming the group to something sexual, such as *"contact named the group My best nude pic's."*, the model was not able to identify it correctly. In these instances, participants often left the group immediately, which was why the conversations were very short. For instance, a 15-year-old female was added to a group chat with porn links, and she described that *"i kept leaving the chat and people kept adding me back in"*. We inspected similar group chats that contained longer messages, where the model was able to identify them correctly.

Many of the false positive instances were short conversations that included automated words from Instagram interactions, such as when a user liked a message on Instagram *"Liked a message"*, send/reply words, *"Shared story"*. These instances were misclassified most likely because they were common in all conversations for both safe and unsafe conversations. Other instances included words from LIWC categories that belonged to unsafe sexual conversations, such as certain swear words. For instance, in a conversation between two friends (16-year-old male participant). This conversation included LIWC categories, such as swear, money, and sexual words that were more often associated with unsafe sexual conversations:

> **Other Person:** *"REMINDS ME OF THAT ONE BITCH IN UR CLASS"..."And honestly I see where ppl come from when they give homeless ppl money, I do. However coming from a family of drug addicts, ppl on and off the streets, ... victims being prostituted; that money goes straight to a pimp where it then went to drugs..."*

This example demonstrated how youth often used profanity and sexualized language when communicating with others, which made true positive cases of sexual risk more difficult to detect and inflates the rate of false positives.

Discussion

In this section, we discuss the key implications from our findings based on our three overarching research questions.

Detecting Sexual Risks in the DMs of Youth (RQ1 & RQ2)

Our classifiers were the first of their kind given our unique dataset and HCML approach; therefore, we provided a baseline from which future works can build upon. We were able to accurately predict real-world private conversations and messages that made youth feel sexually uncomfortable or unsafe on Instagram. Our CNN conversation classifier reached highest performance and was able to identify a higher proportion of unsafe sexual conversations which is necessary in a such sensitive application (recall=0.92). Models with higher AUC had higher recalls for the sexually unsafe conversations, with lower precision as a trade-off. These models had more false-positives than false-negatives, indicating that the models were able to detect most, and in some cases, all, of those that the participants felt uncomfortable. On the other hand, sexual risks message traditional classifiers gained higher precision=0.94 for Sexual class than recall. Next we discuss more about the pros and cons of class specific higher precision vs higher recall.

Precision-Recall Trade-offs

A model with both high precision (the proportion of positive identifications that were actually correct) and high recall (the proportion of actual positives that were identified correctly) would be, unsurprisingly, the most effective solution. However, practically, there are inherent trade-offs between precision and recall for different classes in any given classifier. As such, it is important to consider the context in which a classification system is deployed to understand whether the risk of

false-negatives versus false-positives is higher. We unpack some example scenarios below.

In the case that an algorithm such as ours is embedded in a criminal justice system to identify sexual predators, the legal system typically puts the burden of proof on the prosecution [8] as those who are labeled as sexual predators face criminal charges. Therefore, a model with high precision would align with the goals of the system as it aims to avoid wrongful accusations [55]. However, we advocate that our models instead be used for the purpose of risk prevention (rather than prosecution after-the-fact) and embedded directly within the social media platforms that put youth at-risk, especially with the heightened concern about the well being of youth on Instagram in recent news [9]. This recommendation is consistent with recent U.S. legislation [10] that passed to fight online sex trafficking and stop enabling sex traffickers (FOSTA-SESTA Acts) by making online platforms accountable for user-generated content that promotes sexual violence. To maximize risk prevention, therefore, models with a high recall that aim to prioritize providing support to any potential victims at the cost of false alarms would be suitable for this purpose. Yet, the goals of the primary stakeholders (i.e., adolescents and young adults), in this case, are less clear. Would young social media users get annoyed by too many false warnings of possible unsafe sexual content (similar to what has been learned about overzealous security warnings [162])? Or possibly, would youth benefit from being made aware of the implicit sexual undertones in their conversations (even within their own messages) and benefit from receiving just-in-time support for how to handle such risky situations? Such support may be particularly effective for young adolescents (ages 12-15), as stigma acts as a potential barrier when they seek help [253]. Furthermore, raising risk awareness could help teens identify and appropriately respond to these risky experiences before they escalate to emotional or physical harm [298, 140]. More research is warranted to better understand stake-

[8] https://www.law.cornell.edu/wex/burden_of_proof

[9] https://www.cbs8.com/article/news/health/whistleblower-brings-attention-to-facebook-and-instagram-affecting-young-peoples-mental-health-san-diego-doctor-psychiatry/509-e5d5c810-8491-4186-b6b4-6328ff82fa1f

[10] https://www.vox.com/culture/2018/4/13/17172762/fosta-sesta-backpage-230-internet-freedom

holder needs in the context in which these risk prediction systems are deployed to understand the practical implications of optimizing for precision versus recall.

Conversation-level vs. Message-level Trade-offs

Another computational trade-off is at the unit and/or level of analysis in which classification occurs. While our conversation-level classifier performed the best, message-level classification is necessary for real-time risk detection and mitigation. While risk detection at the conversation-level provides more contextual and linguistic cues for more robust classification, detecting sexual risk or after-the-fact based on entire conservation may come at the risk of being too late. Post-hoc conversational-level detection may be beneficial when attempting to apprehend sexual predators [120], but it does little to protect youth from becoming victimized in the first place [233]. Prior research on Perverted Justice dataset were successful in terms of predicting predatory conversations after-the-fact that risk happens [27, 120, 51], but were not able to detect predatory lines that make a conversation risky [134]. Therefore, we propose a few ways to balance these trade-offs. One option would be to implement sexual risk classification at both levels (conversation and message) for a two-level classification system. The multi-tiered system would more readily be able to support real-time risk detection for mitigation purposes, while the conversation-level detection system could ensure the overall robustness by keeping context intact. Another alternative would be leveraging models that take time dimension into account for predictions that they make, to facilitate more robust risk detection in real time. Having different sexual risk levels introduced a tangible way to allocate resources for treatment, support, and prevention of online risks in an effective manner. The existence of risk in social media could be initially seen as a binary classification task; however, this provides a foundation for future work to study the difference in linguistic styles of risk severity to further aid with providing support to the victims before the high risk levels happen.

Understanding the Private Digital Lives of Youth (RQ3)

Through our human-centered lens, we uncovered important insights about the contextual and linguistic features that were indicative of the sexual risk experiences of youth. Importantly, we found that the age, gender, and the relationship between our participants and others, plus combined features, resulted in the best feature performance compared to more traditional NLP approaches (e.g., Sentiment, TF-IDF). This highlights the importance of including contextual information about people and the relationships between them, rather than relying on linguistic cues alone. We based these contextual features on empirical evidence from prior works [231, 36] that found that gender, age, and relationship context were important factors associated with youth exposure to sexual risks. Past studies have uncovered the importance of relationship context in identifying the riskiness of youth online sexual interactions [82, 235, 277]; yet until now, the computational sexual risk detection has yet to leverage this knowledge in a meaningful way [108]. Additionally, while these prior works were primarily based on the self-reports of youth, our work is the first to triangulate these claims based on social media trace data from youth.

Importance of Interpretability

Our results also highlight the importance of how designing interpretable models using evidence-based research can be superior to black-box approaches. Deep learning models, which mainly prioritize accuracy at the cost of transparency and explainability [226], make it hard for the human users to make sense of *what* and *why* certain features are important. In contrast, our traditional ML models outperformed the deep learning models and added significant value in that they helped us gain deeper insights in the sexual risk experiences of our youth participants. The importance of interpretability of machine learning models have long been advocated by scholars in the HCML area; researchers have argued that interpretability allows us to understand the impacts of machine

learning models on the stakeholders, to think about existing challenges and solutions to making the models' results more human-centered, and to establish the interpretations of what each model does [35, 227]. Comprehending the transparency of the models as to see how the model is based on human-centered approaches [187] is crucial when reviewing the computational risk detection models. Examining each component in the training of the models such as types of algorithms, feature sets, and parameters all contribute to understanding the meaning of the models [126].

The Language of Sexual Victimization

From a more human-centered perspective, our results shed light on the language used to sexually groom, objectify, and victimize youth in private online spaces, which has important implications for both social and computational science, as well as for victim advocacy. In our RQ3 results, we observed that first personal plurals (e.g., "we", "us") were used in safe conversations to show collective purpose and togetherness, while indefinite pronouns (e.g., "it," "that") were more often used by perpetrators of sexual risk to objectify victims and by victims to put distance between themselves and others who made them feel uncomfortable or unsafe online. Further, sexually risky conversations were more likely to contain negative emotions, profanity, words descriptive of biological processes and body parts (e.g., "blood," "pain"), and sex (e.g., "horny," "love," "incest"). A key implication from these findings is that educational and awareness programs for sexual violence and sex trafficking prevention could leverage these insights in training materials that empower women and other vulnerable people (e.g., LGBTQ+) on the linguistic cues indicative of sexually risky dialogue, as well as effective strategies for taking protective measures against these advances when unwanted. A question raised through these insights, however, is whether sexual language used in the formation of *wanted* online romantic relationships [166] mirrors predatory language or is distinguishable from it. Therefore, we recommend that future research also study the language used when youth are forming healthy romantic relationships online to attempt to answer this

unanswered question.

Implications for Design of AI Sexual Risk Detection Systems

Our deeper analysis of contextual feature differences sheds insight in how the thresholds for sexual risk detection algorithms might be optimized for different users. For instance, younger teens were more likely to flag unsafe conversations than young adults; therefore, the classifiers may be fine-tuned to be more sensitive (more tolerant to false-positives) to sexual risks for younger users, while less sensitive (erring toward false-negatives) for young adults. Yet, while social media platforms typically request a users' age upon account creation, the more recent research on age verification [256] should be considered given youth are known to sometimes lie about their age when joining these platforms [117]. Similarly, algorithmic sexual risk detection systems could identify the nature of the relationship to make sexual risk detection more attuned to conversations between strangers and acquaintances, as opposed to friends, significant others, and family members. This would be an alternative to the stricter decision of Instagram and other social media platforms to ban strangers from having direct message conversations with minors who do not follow them [11]. Of course, such deploying AI in this way would require extensive user evaluation and design work to make sure that misclassifications based on age and relationship type did not unintentionally burden or harm end users in unexpected ways.

Integrating an ML system to automatically flag private conversations as sexually risk or not should be done with great care for social media users. In order to preserve users' privacy, light resource consuming local pre-trained models [170] could be implemented to detect online risks for youth, not sending all the private information to the cloud to be continuously monitored. We have seen how difficult the problem is when big platforms struggle to solve similar problems of harmful be-

[11] https://www.deseret.com/opinion/2021/3/24/22348616/is-tiktok-facebook-twitter-safe-for-kids-privacy-settings

haviour (e.g. hate speech, disinformation) on public discourse [12]. In a private setting, the problem is even more complex since there may be many subtleties in the communication between two individuals that would render the problem ill-defined for an ML system. Therefore, it becomes more important on what decisions will be made when such instance is detected by the system. For instance, when a sexually risky instance is detected by the ML model, it should provide a suggestion to the youth user and the user should be able to provide feedback to the model and make the final decision. Using such human-in-the-loop approaches [306] would help the user be in the control of the uncertainties of the real world. In addition, identification of sexually-risky content could get conflated with content that mentions sex in general such that in cases people discussing uncertainties around sexual identity, e.g., might be disproportionately harmed. For having scalable solutions in real world, handling possibility of false-positive/negative instances becomes of an important design decision. Since unnecessary suspicion on innocent conversations or missing instances of harassment can be unacceptable in the real-world.

Limitations and Future Work

A key strength but also a limitation to the generalizability of our work is that we collected a difficult to obtain dataset of private Instagram conversations from youth (ages 13-21). Therefore, our results should only be generalized to this population. Further, since our analysis was based on Instagram, our results may be constrained by the unique affordances [284] of Instagram and not generalizable to other social media platforms. Therefore, more research is warranted on examining private conversations that occur on other platforms with other user groups to validate whether our sexual risk classifiers are transferable to those new contexts. It would be interesting to investigate how the effect of contextual features (e.g. age, gender, and relationship) would change if the

[12]https://time.com/5855733/social-media-platforms-claim-moderation-will-reduce-harassment-disinformation-and-conspiracies-it-wont/

age window for the participants were widened. Also, we focus solely on textual and contextual information, rather than media (e.g., images, videos, links), which has been the subject of inquiry of prior work in the HCI community [12, 266]. In our future work, our aim is to consider multi-modal approaches for sexual risk detection that includes both textual content and media.

While our results are promising for detecting unsafe sexual conversations experienced by youth, we faced the challenge of imbalance in our dataset, having relatively smaller unsafe interactions than safe ones. For creating a balanced dataset for the unsafe sexual conversation classifier we used under-sampling to reduce the number of safe conversations. The main issue with under-sampling is the possibility of losing informative instances from the majority class while deleting the instances. To make sure that the sampled examples were diverse enough, we manually checked the remaining samples. Yet, before our models are deployed in large, user-based platforms, more unsafe training samples are needed to reduce false-positives. To address these limitations, we are in the process of collecting more data. As a future work, we also plan to have youth evaluate the quality of our sexual risk classifiers by designing and deploying a web-based risk detection system where they can upload their Instagram data for the system to identify risky content. This will complete the HCML loop of having users direct feedback on the performance of our algorithms, so that they can be further refined for future real-world use and impact.

Conclusion

The core contribution of this work is that our findings are grounded in the voices of youth who experienced online sexual risks and were brave enough to share these experiences with us. To the best of our knowledge, this is the first work that analyzes machine learning approaches on private social media conversations of youth to detect unsafe sexual conversations. In addition, this work highlights the importance of contextual and implicit features on identifications of unsafe sexual

conversations and provides a good indication of how different methods and features perform when addressing this problem. Given the wealth of data we have collected, but have yet to analyze, we welcome other HCI and ML researchers to join us in our efforts to use HCML as a proactive way to protect and empower youth online.

CHAPTER 6: STUDY 4: REVISITING UNCOMFORTABLE ONLINE INTERACTIONS: UNDERSTANDING THE IMPACTS ON PARTICIPANTS AND RESEARCHERS FLAGGING SOCIAL MEDIA PRIVATE CONVERSATIONS FOR UNSAFE OR UNCOMFORTABLE CONTENTS

Abstract

Human-Computer Interaction (HCI) research values the viewpoints and data of end-users, but often this information can be very sensitive and/or triggering. Therefore, it is important to reflect on how HCI research participants view their participation in such research through a case study of asking youth to donate their Instagram data and annotate their private message conversations for their past online risk experiences. We conducted 29 interviews with youth (ages 13-21) who donated and annotated their Instagram data for online risks. The goal is to utilize this ecologically valid dataset to provide insights from teens' private social media interactions. Additionally, we interviewed 12 research assistants who annotated the participants' private conversations for online risks to investigate how it affected their mental health and perspectives. We found that youth were comfortable sharing their data to advance the adolescents online safety research as long as their data be kept private, not be used for making profit, nor be shared publicly. Also, reviewing their conversations and flagging them for risks taught them valuable lessons including insights about their communications with others, how they respond in uncomfortable situations, and their online habits. Although the increased risk awareness came with the price of awakening some negative emotions such as feeling uncomfortable of the past inappropriate interactions. In a few cases, participants expressed stronger emotions such as nervousness and getting upset. Our results

demonstrated how context and connotations of conversations matter when it comes to risk-flagging such that participants considered the relationship and intents of people involved in conversations, and their emotional feelings. We discovered that research assistants were surprised to observe the frequency and variety of the risks that youth are exposed to, but it did not cause them emotional distress. They learned more about the online risks and privacy issues and how they could keep themselves and their loved ones safe online. Our research provides insights and implications for design for sensitive research with participants and their data annotation for online risks and mental health of researchers.

Introduction

It is important to engage people in Human-Computer Interaction (HCI) research, especially as it relates to their lived experiences. HCI research has moved to increasingly studying vulnerable populations and online risks as a way to mitigate them and empower end users [181, 231, 18]. However, in doing this, it could potentially cause unintentional harm and/or trauma to both research participants and the researchers [17]. Since reviewing sensitive data for flagging for online risks inevitably might bring up memories, particularly for the ones who experienced trauma. HCI community [201, 62, 289] discussed the concerns about how to protect participants' and researchers' emotional wellbeing in sensitive research such as designing online support systems for bereavement [181], sexual abuse [118], stigmatized experiences such as pregnancy loss [18], and mental health challenges [215]. As Artificial Intelligence (AI) and big data-driven approaches push to address the issues related to online safety of youth; yet much needs to be uncovered related to their prospective of unsafe interactions and how it would affect them to flag their unsafe private messages for ground truth. The potential emotional risk to participants and researchers is unknown in terms of impact. It is not feasible to predict the content which might be trauma triggering to

participants/researchers as it is unique to individuals. Therefore, as researchers, we must conduct meta-level research to understand how both our research participants and research assistants are impacted and reflect upon engaging in high-risk, high-reward research. In social sciences research, 'Trauma-informed' approaches in research design and practice is well-established; however, it has not yet been applied to HCI researchers. These approaches can build upon existing ethical and methodological frameworks to inform how HCI researchers conduct research with participant and consideration for the mental health of stake holders in research. As suggested by heuristic guidelines for research with vulnerable populations by Walker et al. [287], it is important to perform a post-research to help the community or the targeted research population. In this work, we conducted a meta research on the ethical treatment of research participants who engage in more than minimal risk research and researchers who annotated their data. To better understand the affect of reviewing and labeling unsafe/risky content of youth and research assistants, we pose the following research questions:

- **RQ1:** *How participation in an adolescent online safety research that asked youth to donate their social media data and flag their private messages that made them feel uncomfortable or unsafe impacted participants in terms of potential harms and benefit?*
- **RQ2:** *What are the potential positive and negative impact of data flagging on research assistants who were asked to review the private messages of youth to annotate online risk experiences?*

To answer the research questions, we conducted retrospective semi-structured interviews which helped us evaluate a data collection and annotation for adolescents' online safety. We qualitatively analyzed 29 participants and 12 research assistants who annotated participant's social media private conversations for online risks. Our research contributes to HCI and adolescent online safety research by providing insights on how participation in more than minimal risk study and sharing

private sensitive social media data and flagging it for online risks impacted youth. Complementarily, retrospective semi-structured interviews with the third party annotators of the data helped us to understand their thoughts during the annotation process and its effect on their wellbeing. We provide recommendations for research ethics for both participants and researchers working in the sensitive area of adolescents online safety. We provided design implications for sensitive research to ensure the well-being of the participants and researchers. This paves the way for reducing the negative impact of viewing and annotating unsafe online interactions and for protecting them accordingly.

Background

In this section, position our work with the relevant literature in the area of wellbeing and ethical considerations of sensitive research with participants and researchers.

Participants' Wellbeing and Research Ethics

Across all disciplines in the United States, Belmont Report (finalized in 1979) has widely used in the USA to protect the rights of all research subjects or participants. There are three main components: 1) respect for persons, 2) beneficence (minimizing harm), and 3) justice. Institutional Review Boards (IRB) prompt researchers to go through a process to document potential risks to the participants and plan to prevent and mitigate potential harm to them based on the main components in Belmont Report. Although the core principals of the report will continue to protect the rights of human subjects, the report is now more than 40 years old. Research related to new online technologies and social media platforms poses new ethical considerations, which raises challenges to IRBs and researchers [202]. Therefore the Belmont Report does not account for all the emerging

issues about research particularly on social media and there has been criticisms to it [101].

In addition, the institutional governance procedures to ensure the core ethical principles are known as "procedural ethics" [113] in which does not necessarily include the unanticipated "ethics in practice" issues which is different in context of each type of research project. For instance in qualitative interviews, researchers emphasised how it is important to establish a rapport with participants and make a safe trusting environment for interacting with participants [212]. Besides, Various types of research settings impose diffident challenges, such that having participants feel safe and comfortable in a remote online study includes different challenges.

Researchers investigated the ethical challenges of using public social media for research and best practices of utilizing public data for research [94, 294, 216]. For instance, Fiesler and Proferes [94] conducted a survey of Twitter users' perceptions of the use of tweets in research. They found that the majority of participants felt that researchers should not be able to use tweets without consent, though it depends on the context. Although, the participants' view on how their private data is going to be used and their concerns is an understudied area.

Research Ethics of working with Minors and their Online Safety

Researchers [287, 25, 29, 229] investigated ethical challenges especially working with vulnerable populations such as adolescents. Walker et al. [287] provided heuristic guidelines to consider the needs of vulnerable and marginalized populations when in the research process. This heuristics for research design with vulnerable populations includes considerations for pre-research (e.g. need findings of vulnerable community, relationship of researchers with the community, appropriate compensation, consent), during the act of research (e.g., power differentials, data considerations, empowering participants), and post-research(e.g., presenting findings, positionality of researchers, dissemination of the results). Antle et al. [25] stated ethical challenges and considerations while

conducting research with children living in poverty in Nepal. Only a few researchers [29] investigated participants review regarding online safety research. For instance, Badillo-urquiola et al. [29] had 20 youth co-design adolescent online safety research and interviewed 13 of their parents. They found that adolescents were motivated to share their data to benefit society, while they feared getting in trouble. Whereas parents wanted researchers to solve problems particular for their teens. Our works contributes to the literature by providing insights into the context of conducting sensitive research with youth about their private unsafe interactions on social media.

Researchers' Welfare in Sensitive topics

IRB or traditional forms of governance for research, mostly do not consider questions about research ethics and protection of researchers especially for sensitive research. There is a growing body of research acknowledging the sensitivity of conducting research that involves exposure to traumatic information. Specifically, previous research has investigated the potential adverse impact of being involved in such research on the researchers' mental well-being. For example, Mckenzie et al. [186] interviewed eight research assistants who collected and worked on self-harm and suicide attempts clinical notes. They found that research assistants experienced a wide range of challenges when undertaking such research, which includes being emotionally or psychologically unprepared for the level of detail and the sensitivity of the information in the records, being personally drawn into individual stories, and feeling emotional exhaustion from the cumulative effect of processing the data over a long period of time. Given the sensitive nature of the data, conducting such research warrants coping strategies to ensure the emotional and mental safety of researchers [295, 76]. Williamson et al. [295] conducted an interview with 10 researchers and presented their coping strategies to minimize the adverse emotional impacts when working on Gender-Based Violence (GBV) research. They reported coping practices such as having a distraction time to read a magazine, watching TV, exercising, or going to therapeutic counseling. Such practices might be

useful for short-term relief, which could not fully mitigate the harmful mental effect. Therefore, adopting a trauma-informed care lens when conducting research has become a critical consideration to help researchers to maintain the professional distance and minimize the adverse mental health impacts on them [76, 186].

Researching on sensitive topics requires well-documented and designed guidelines for the researchers to protect the researchers' mental health similar to the health professionals [75]. For instance, Vidgen et al. [281] shared their developed guidelines for researchers to minimize the harmful mental effects that might be caused by the cumulative exposure to viewing and annotating online abusive content [281]. They provided guidelines to run the project effectively, create a supportive work environment, and implement responsible work practices. Although these guidelines might be useful to mitigate the impacts of working on such content, these guidelines were researchers' efforts that might not be applicable to other type of content such as self-harm records or other projects. Therefore, there has been a high demand for universities, granting institutions, and academic leaders to expand the research ethical considerations to protect the researchers and assess the potential risks on their mental well-being [17]. In addition, there have not been a study specifically for the perspectives of research assistants and their mental health when annotating youth's social media conversations for online risks.

Methods

Below, we provide an overview of our study, describe our methods, provide details regarding our data analysis approach.

Table 6.1: Annotators' Codebook

Dimension	**Code**	**Definition**
Surprised	Frequency	How frequent online risks are, lots of spam, sexual risks, etc.
	Personal	Personal conversations and debates.
	Escalated	How risks were evolved/escalated with sending more videos links.
Concerns	Reporting	what if they found something bad that they had to report.
	Distress	Had negative emotions.
	Uncomfortable	Felt bad or uncomfortable in the moment but did not affect them.
Learned	GT	Ground-truth importance for ML
	Advice	Give advice to family/friends and watch out more.
	Privacy	Became more privacy aware.
	Reflection	Made personal reflections.
	Risk	Gained online risks knowledge.
	Positive	Had positive impact on their mental health.
Support	Help	Be there to help, guide, clarify, define risks.
	Strategy	Tell them not to take it personally! Encourage them to take breaks.
	Motivate	Give them motivation, and emphasize importance of their work
	Work	Have workshops and assign a smaller number of conversations.
	Tool	Improve annotation tool.

Study Overview

We built an online system for youth (ages 13-21) to donate and annotate their Instagram data integrated with their self-reported data. In this project, Instagram Data Donation (IGDD) project [234], the goal is to improve adolescent online safety by creating ecologically valid dataset for training machine learning models for adolescents online risk detection. 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 a paper [234], we explain our design and study decisions, lessons learned through the design, development, and the data collection process.

Table 6.2: Participants' Codebook

Themes	**Code**	**Definition**
Motivation for Participation	Incentive	Motivated to participate by receiving gift card incentive.
	Interest	Motivated to participate by their interest to social media and how it works.
	Research	Motivated to participate contributing to research.
Data Use Envisions and Concerns	Privacy	They envisioned their data is kept private and confidential.
	Purpose	They envisioned their data is being used for purpose of research.
	Selling	They envisioned their data is not to be sold to third parties or for profit.
Flagging Emotions	Positive	Felt only positive emotions while flagging past messages.
	Negative	Felt only negative emotions while flagging past messages.
	Mixed	Felt positive and negative emotions while flagging past messages.
	Neutral	Felt neutral while flagging past messages.
Flagging Risky Content Criteria	Context	Their flagging criteria was based on the context, connotations, tones of conversations.
	Real-life	Their flagging criteria was based on whether the interaction would influence the participant's life.
	Relationship	Their flagging criteria was based on relationship they had.
	Intent	Their flagging criteria was based on the intent of people.
	Emotional	Their flagging criteria was based on how messages made them feel emotionally.
Discomfort during FlaggingPast Conversations	Uncomfortable	Not positive to look back at their past unsafe interactions, made them feel uncomfortable.
	Upset	Felt very sad and emotional while looking at past messages.
	Angry	Reacted angrily or annoyed while looking at past messages.
	Weird	Had a strange/weird feeling looking at past messages.
Positive Impacts of Flagging Past Conversations	Usage Change	Flagging their past conversations made them change their habits and how they use social media.
	Reflection	Flagging their past conversations made them reflect more on people's intentions and their respond.
	Same	They learned noting new from flagging their past conversations.
	Communications	Flagging their past conversations helped them change their communications with others.
Reason Discontinued	Device	The reason they discontinued was that they didn't have the right device (laptop or desktop).
	Technical	The reason they discontinued was that they had technical/upload issues.
	Privacy	The reason they discontinued was that they had privacy concerns.

Interview Study Design

We performed a retrospective semi-structured interviews with the youth that participated in the IGDD study to understand how they felt about reviewing and flagging their past risk experience. The goal is to have them reflect on their past experience to evaluate how they felt about this interface and provide implications for design and best mental health practices for protecting them accordingly. We designed a semi-structured interview script based on how they felt participating in the study, reviewing unsafe/risky conversations, and how that affected them. We asked follow-up questions to clarify interesting discussion points that came up during the conversations.

Interviews Data Collection and Recruitment

We invited eligible participants who started/completed IGDD study. The participants over 18 years old were required to fill out an adult consent form while participant under 18 years old required parental consent and teens' assent before participating. We conducted interviews over a 30 minutes Zoom session. We had a risk mitigation plan in place to ensure the safety of the participants. To ensure the protection participants in case of any off-camera interaction, we also took some safety measures:

- The participants were informed beforehand via consent forms and reminded before the Zoom session that keeping the audio and video turned on at all times during the session is mandatory for participation.
- We used Zoom's feature to request permission from participants to unmute them during the call. If participants allow, this feature allows the host to unmute participants at any time they are muted.
- We used the Zoom security feature to disable breakout rooms to avoid any off-camera interaction.
- Using the Zoom security features, we also disabled participant's ability to record or save any recordings locally. Participants will not be able to override the hosts ability to record the Zoom call.
- We used the Zoom security feature to disable private chats between participants.
- Additionally, researchers shared their contact information and help resources with the participants for them to reach out at any time before, during or after the study.

We incentivized participation with a $20 Amazon gift card distributed to the participant via email upon completion of the interview. We conducted a total of 29 interviews with youth participants from IGDD and a total of 12 with research assistants who annotated the youth data and agreed to participate. From the 29 participants, 21 participants completed the IGDD study and passed the eligibility requirement, but 8 of them did not continue the IGDD to upload and flag their Instagram data. The reason for having the participants who discontinued Phase 1 was to understand why they did not continue the study.

Qualitative Data Analysis Approach

We used content analysis [88] and thematic analysis [54] as our qualitative approaches to analyze the interview data. The content analysis was used to understand the convergence and variance across the participants' responses for each question. The first author coded all the interview transcripts. The coauthors reviewed the consistency of the codes iteratively throughout the data analysis phase. The codebooks for annotators and participants are summarized in order in Table 6.1 and Table 6.2. After the structured content analysis, we conducted a thematic analysis of emergent themes.

Results

We present our findings by first discussing characteristics of the participants' profiles. Then we discuss the major themes that emerged from participants' and annotators' interviews.

Table 6.3: Participants Demographics

ID	Type	Age	Gender	Race
P1	Passed	19	Female	White/Caucasian,Prefer to Self-Identify
P2	Passed	18	Male	White/Caucasian
P3	Passed	15	Gender-fluid	Black/African-American
P4	Discontinued	18	Non-Binary	White/Caucasian
P5	Passed	18	Non-Binary	Asian or Pacific Islander
p6	Passed	18	Female	White/Caucasian,Black/African-American,Hispanic/Latino
P7	Discontinued	18	Male	-
P8	Discontinued	21	Non-Binary	White/Caucasian,Asian or Pacific Islander
P9	Discontinued	21	Female	Asian or Pacific Islander
P10	Passed	14	Female	Asian or Pacific Islander
P11	Passed	17	Male	White/Caucasian
P12	Passed	16	Female	Black/African-American
P13	Discontinued	19	Female	Asian or Pacific Islander
P14	Passed	15	Female	White/Caucasian
P15	Passed	18	Male	Black/African-American
P16	Passed	17	Female	White/Caucasian,Black/African-American
P17	Passed	18	Female	White/Caucasian,Asian or Pacific Islander
P18	Passed	21	Female	White/Caucasian
P19	Passed	16	Female	White/Caucasian
P20	Passed	14	Female	White/Caucasian
P21	Discontinued	20	Female	Black/African-American
P22	Passed	14	Female	Black/African-American,American Indian/Alaska Native
P23	Discontinued	17	Female	White/Caucasian,Asian or Pacific Islander
P24	Discontinued	16	Male	White/Caucasian,Hispanic/Latino
P25	Passed	15	Female	Black/African-American
P26	Passed	17	Male	White/Caucasian,Hispanic/Latino
P27	Passed	14	Female	Black/African-American,American Indian/Alaska Native
P28	Passed	16	Male	White/Caucasian
P29	Passed	17	Female	Hispanic/Latino

Participants' Profiles

Table 6.3 presents the demographics of the participants. Of the total of 29 participants, 13 of them were over 18 years old and 16 were under 18 years old. Twenty-one of them completed the study and eight of them discontinued the study. Most of our participants were female (N=18, 62%) with seven (24%) who identified as male and four (13%) as non-binary. Eight participants identified as

white or Caucasian;other participants identified as black/African American (N=5), Asian or Pacific Islander (N=3), multi-ethnic (N=11), and one participant preferred not to answer.

Participants Findings

We present the findings from participants' interviews in this section based on the major themes emerged.

Comfortable Sharing Their Data for Research Purposes

Participants mostly (N=15, 52%) participated in our study motivated by contributing to the research on adolescent online safety or they were interested in the subject (N=11, 38%). Although, some participants were motivated to receive the incentive (N=8, 28%) and for personal reasons. They expressed how important it is to solve the online safety issues for adolescents and how they would be happy to contribute to research in this area. For instance, the following participant, stated that although it look substantial amount of time to complete the study, but contributing to the research was worth the time and effort:

> *"I'm glad to participate but took longer than expected, but I believe it was worth it. It was worth it because it made it possible for my data to be used for research."* P2

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. Based on consent/assent forms statements, most participants (N=20, 69%) envisioned their data being used only for research purposes and

providing insights for solving problems. Participants did not express much of a concern, as long as their de-identified data being used for research purposes.

> *"I felt comfortable sharing my data with a large university and I never felt my data was compromised. I imagined my data would be used towards discussions about social media and the trends occurring among students and teenagers and how social media can affect mental health and messaging on social media specifically."* P3

Participants cared about how their data could be used to create online safety solutions to prevent the risky interactions for benefiting youth.

> *"I understand data will be used to help prevent risky messages to figure out some sort of technology to prevent. I don't regret sharing my data. Just don't use it against me."* P10

Almost half of participants (N=14, 48%) expressed how important it is to keep their data private. They mentioned they want their data to be de-identified and they do not want any personal information of them being disseminated outside of the research team or be shared publicly:

> *"I felt more comfortable sharing my data… this could be potentially helping. I can't think of any concerns right now; as long as my name isn't plastered everywhere."* P6

A few participants were a little concerned and self cognizant on someone else like researchers are going to look at their messages.

> *"The fact that, other people, besides me are like looking at these conversation I had with like someone else could be like the privacy thing. I was a little concern participating in the study."* P29

They (N=6, 21%) did not want their data to be investigated individually and wanted the data to be used to make aggregated insights.

> *"I envisioned the data being used for mass data analysis. I do not want my data to be used for very individualized stuff. I do not want my data to be shared outside the group of researchers."* P5

For instance, they mentioned how they want their data to be categorized into groups to provide summarized insights:

> *"I Would hope the data conversations are being summarized instead of individualized. For the most part, no concerns about how the data is being used."* P9

They (N=4, 14%) also mentioned that they do not want their data to be sold to third parties, to be used for advertisement, or be used for making a profit.

> *"I would not want data to be used for profit purposes. I feel researchers need to be careful of the data."* P12

Overall, youth were willing to contribute to advancing the adolescent online safety research, as long as their data be kept private and unidentifiable, and not be used to against them to make a profit.

Reflecting on Past Messages Increased Their Awareness (some discomfort)

Reviewing and flagging their past conversations, participants (N=11, 52%) changed the way they communicate with others and helped them improve it. They thought about how their friendships

were, how they trusted people and shared their information with others. For instance, P1 talked about how it was a beneficial experience to look at the relationships she had and the way they communicated:

> *"I think it was probably a beneficial experience just in terms of scrolling and evaluating past relationships in terms of looking at the other people like oh Why did I hang out with this person like how do I react fast, but also, it was kind of it was like a weird feeling, but I think. It's largely beneficial to see things that I had sent like a couple of years ago, not necessarily because they were explicitly bad, but just like cuz the way that I communicate, I think it has changed was just interesting to see the difference."* P1

It also changed the way they use social media and their online habits (N=11, 52%). Frequently, participants mentioned how they use to respond to every message, but after reviewing those interactions they learned that they can stop responding and ignore more messages. For instance, P10 mentioned how they realized they can used features such as reporting and blocking more often rather than responding to those uncomfortable/unsafe situations.

> *"It wasn't too negative an experience, just realizing what I should do in these types of situations in the future. My perspective of a few of the conversations changed after reviewing them. I realized I should be better about reporting and blocking risky messages immediately instead of responding."* P10

After flagging their conversations, they reflected more on the dynamics of interactions online and people's intentions and how they respond to them (N=8, 38%). Although, for some (N=8, 38%) participants their online behavior stayed the same and did not change. P6 expressed how the study helped her recall situations online that she made mistakes and now she is more self-aware:

"I think the study helped me remember some things and I think "that was kinda messed up"; I am now more self-aware." P6

From the participants who completed the study and were interviewed (N=21), while flagging messages, described variety of feelings with various degrees to which they experienced these feelings. To categorize these emotions more broadly, some of had negative emotions (N=8, 38%), mixed feelings (N=7, 34%), and some (N=6, 29%) felt neutral and did not have any feeling while flagging messages. These feeling included feeling uncomfortable for almost most half of the participants (N=9, 43%). For example, P5 felt uncomfortable looking back at the conversations since they realized that they kept engaging instead of ignoring in conversations that they did not feel comfortable.

"Reflecting on messages made me feel uncomfortable because while I was rereading the messages, there were instances that I felt uncomfortable at the time, but I kept engaging in the conversation instead of ignoring the message sender. I have negative feelings looking back at my messages, nothing could be done on your side. I have learned that negative feelings towards a conversation are valid, and should just ignore those messages instead of continuing to talk; learned to reflect more before continuing conversation." P5

P10 described her feelings with a little regret but not too negative, since she wished she could change some things she said, but now she knows better how to react on those types of situations:

"I felt a little regret reviewing these past messages, wish I could have taken back the things I said. It wasn't too negative an experience, just realizing what I should do in these types of situations in the future. My perspective of a few of the conversa-

tions changed after reviewing them. I realized I should be better about reporting and blocking risky messages immediately instead of responding." P10

Some participants (N=5, 24%) emotions were stronger than feeling uncomfortable, reviewing the unsafe interactions made them upset/sad and had a negative impact on them.

"Reflecting on past messages made me nervous and worry; even though they were a long time ago, they could still have an impact on me in the future. It feels like reading your own messages and reflecting on yourself could have an impact on your mental health. It makes you question some of the things you had said in the past." P29

Unfortunately, a few participants felt angry (N=5, 24%), because reflecting on their unpleasant interactions made them rebuild situations in which they felt powerless:

"Made me feel uncomfortable, angry, and powerless, not supper pleasant to go through, it happened a long time ago and it brought up stuff and didn't make me feel that great." P11

A few (N=4, 19%) participants mentioned they felt weird reviewing their past interactions. Those feelings were mostly because of the nature of the interactions that happened and participants mentioned that there was nothing that we could do better in order to alleviate those feelings.

Context Really Mattered When It Came to Risk-Flagging (relationship type, intent, etc.)

For flagging conversations for risks, participants criteria were the intent of the person who they are talking to (N=14 , 48%). They mentioned that it was sometimes hard to know the intention of the people they do not know.

"My criteria was if I could determine person' intention was not good." P23

The emotional feelings that they had while they were reading the messages (N=14 , 48%) and what affected them in real-life was also a prominent factor:

"If I felt anxiety or negative emotion come up when I read it, then marked it as unsafe. Being in a situation in the conversations, the context of what was happening in real life, what is going to happen physically in life - all mattered in flagging risky messages." P18

Participants (N=8 , 28%) mentioned that the relationship type of the people involved in the conversation also affected on the riskiness of the messages. Participants mentioned that they can easily ignore messages from strangers, but it is harder to not engage with people they know. Although some mentioned that they used to engage with strangers and that would be of a high risk for them since they had shared information with people they did not know or trust.

"Relationship with sender influenced the criteria of flagging messages. It is easier to not engage with strangers. I felt more pressure to respond to those that the I was acquaintances/friends with. My perspective has been changed since I was younger because I'm a bit more cautious now and was unaware of sexual messages when I was younger." P5

Lower risk level conversations usually involved unwanted messages from strangers or bots that the youth did not respond to. While higher risk level conversations included more explicit content that affected them in real life or has been repeated.

"Safe/unsafe messages depends on the context. For example, A friend threatened to tell my mom; made me very uncomfortable because it was intruding on my private

life. I feel harassment social media messages as medium and high-risk level which are anything that intruding in my real life and I see the person daily in school (someone I know)." P12

Participants (N=5 , 17%) emphasized how important the context and connotations behind the conversations are for determining the riskiness. For instance, if what happened online affected them in real life mentally or physically or not or vice versa was a criteria for flagging unsafe interactions:

"I have decided what was safe and unsafe by connotation. A negative connotation (whether it was something that offended me or if I was arguing with the person) will be unsafe. I have only really experienced people sliding into my DMs like "hey baby" or stuff like that; generally, I block them." P28

Overall participants took many factors into considerations when determining unsafe interactions. Most importantly, their criteria was based on the intent of the people involved and the relationships they had with them. Also how it affected their life especially emotionally and the context of the conversations.

Most Who Discontinued Did So due to Technical Difficulties

Majority (4 out of 6) of the participants who discontinued the study did so because of upload or technical issues. They could complete the first part of the study on any device such as a mobile device, but in order to upload their Instagram data they needed a desktop computer. The reason is the limited space on mobile devices and difficulty to store the file, browse, and upload it. For instance, P4 mentioned how they could not complete the study because other than a Chrome book from their school, they did not have a proper device:

> *"I did not complete study Phone space was limited Have Chrome book from school; the site was blocked on my computer. I do' not have the technological availability to complete the study."* P4

Only two participants who discontinued the study expressed privacy concerns related to sharing their private messages:

> *"To relinquish private messages in one quick data file, I felt especially uncomfortable. I want to keep that information private; I'm uncomfortable to bargain with that topic. I did not complete the study for that reason."* P7

Overall, most of the participants who did not complete the study had technical difficulties not privacy issues. The researchers needed to work on minimizing the technical difficulties. For the a few participants with privacy concerns, other options such as upload partial messages should be provided.

Annotators Findings

In this section, we provide the results for interviews with research assistants who annotated the data.

Surprised to See The Types of Risks Teens (especially young teens) are Exposed to, but Did Not Cause Emotional Distress

Most (N=7, 58%) of the annotators in our project were surprised to see how frequent online risks are happening to youth, but that did not cause emotional distress to them. Some (N=4, 33%) of

them mentioned that some messages or media made them uncomfortable. Although that feeling was only temporary and did not have a long-term or major impact on them.

> *"I guess a lot of the creeps was like guys messaging girls. How frequent it was and the things they would say would surprised me. That stuff that was like really gross me out in the moment. Nothing like that hurt me like mentally or physically that I carried on."* RA10

Mostly, annotators felt uncomfortable looking at sexually suggestive media. Since our study involved minors, there was the potential for identifying child abuse and neglect, the research team members were mandated reported and needed to report possible child abuse/neglect. In a few cases (N=2, 17%), annotators were worried if they find something that they have to report such as child pornography.

> *"Sometimes I guess when people send images that were kind of graphic memes that would make me uncomfortable!"* RA8

Although, less frequently (N=2, 17%) annotators mentioned that they were sad to see all the unsafe interactions that youth deal with:

> *"People getting so many bad messages from strangers that affect them negatively. It breaks my heart. Really important to protect these people that cannot say no.Something that I don't want people experience in their life. Really sad that it was happening so frequently with so many people. "*R6

Though, the sad feelings were only short negative feelings that disappeared after a short amount of time:

> *"The only messages that were triggering only lingered in my mind for about 10 minutes; mostly related to sexuality. Bothered by messages that insulted or judged someone based off of their sexuality. Sad to see that kids are growing up believing these things"* R11

Overall, annotators were surprised by how frequent online risks happen to teens and at times it was unpleasant to review those interactions, but it only affected them at the times they were reviewing those messages and they did not carry the negative feelings with them.

Learned more about Online Risks made them Reflect on their Past Experience, Privacy, and Safety

Most of the annotators reflected (N=7, 58%) on how the annotation task affected their perspectives on social media interactions. They reflected on how social media is used for friendships, online social dynamics, and social connections. Also, they reflected on how some people are more vulnerable to online risks. For instance, R7 reflected on how people have this perception of having more friends online is good, but they do not think as much on how adding people that you do not know could be dangerous:

> *"I thought more about how risky it is for other people. Sometimes there is a sense that you have more followers the better, then you accept more people, but you don't really know them and they are looking at your content and start making your friends their friends,.. it is dangerous sometimes."* R7

The annotation task not only made them to think about the situations where they/or other people were vulnerable, but also made them reflect if they ever hurt someone:

"Thought of some of the ways that I might have hurt someone either on social media or in person" R10

Annotating social media conversations of youth for online safety made half (N=6, 50%) of the research assistants to become more privacy aware. Some of them limited the use of social media or totally removed their accounts from their lives. They became more cognizant about who their friends and connections are and removed the people that they do not know:

"I became more aware of who is on my social media. I removed anybody that I didn't know." R11

They explained the reasons behind their decision of limiting/removing social media including amount of unsafe interactions that happen online especially sexual risks, and amount of time that gets wasted online.

"Deleted social media. I saw all the things that happened, and how much time it takes" R10

Many of annotators (N=6,, 50%) expressed the things they learned about online risks and gained more experience on types of risks that are happening.

"Reflected more on these types of messages, rather than the serious unsafe messages." R1

The task made them more cautious of online safety especially about younger friends and family. They more often (N=4, 33%) gave advice to their younger siblings, relatives, or friends.

> *"I was among first people on messaging apps and I experienced in the first time. More innocent at the time. Dialogue online was very useful and safe at the time, now it's become dangerous. It was cool, it worried me about younger sister to stay away!"* R9

Although, they provided ways to their friends and family on how to be safer online by implementing more privacy measures:

> *"I give friends advice on how to avoid online risk; suggested them private account, only allow friends to follow you, block messages from strangers, do not open any links or photos"* R6

A few (N=3) annotators mentioned that the annotations had positive impact on them when they reflected on their past experiences. They also felt they have positive impact by helping to detect online risks for youth. For instance, R2 felt grateful that she did not use social media to be exposed to a lot of risks, especially when she was younger.

> *"Made me reflect on how I felt the same way that these teen participants did struggling in their mental health, when I was a teen, and now life is much better. Somewhat a positive experience."* R2

Overall, annotators gained more insights on online risks which make them reflect more on their own experiences and made them give advice to people they care about.

Needed More Context Provided by Annotation Tool

Most of annotators (N=9) mentioned that annotation tool should provide more information including remaining participants, percentage of messages remaining in each conversation, metadata,

relationship between people involved. etc. Some of the conversations were difficult and confusing to annotate without knowing the story and information around them. Oftentimes, there were multiple people involved in those conversations and the underlying context were not available So, some of the suggestions to improve the annotation tool include:

> *"Something to add is to be able to see what persons were in multiple conversations (group chats included) and look at each of them as a set; Add a button to flag a message that the annotator wants a second opinion on by another annotator."* R5

Suggestions included making the annotation tool more reliable and to add a few minor features.

> *"if the annotation tool could highlight the words that are bad that can help the annotators."* R6

As some conversations were long, having more features to help annotators easier identify unsafe interactions such as a feature to highlight risky interactions are useful.

Research teams should Support them

Annotators felt that most of their feedback were constantly addressed. They (N=6) found the support and help from their mentors and the team very important. They mentioned that some conversations were tricky and unclear to annotate, but the team were always there to consult on confusing cases. They provided invaluable feedback to improve the process. They found motivations and encouragements from the mentors very important:

> *"it's okay if you need a break from annotation tool. Expose annotators as to why we're doing what we do and what the bigger picture is to motivate annotators"* R2

For instance, R6 is suggested that the mentors provide strategies to annotators to not take messages personally, provide opportunities for them to work together:

> *"Make sure annotators are not taking those messages personally or close to their heart .Give a chance to annotator and give it to another annotator, have small workshops once a month and tell them what you are doing is good and encourage people. How much value it has for."* R6

Broadly, the support and tools that the mentors provide for annotators are of a crucial importance which can alleviate some of the negative experiences resulted from the nature of reviewing unpleasant online interactions.

Helped Teach Computer Science Undergrads the Importance of Ground Truth in Machine Learning

Through the annotation task, half of research assistants (N=6, 50%) mentioned how ground truth labels are important for training supervised machine learning models.

> *"Annotating participant's data, flagging risk level and category is important because it provides the baseline for the model and quality of the annotation will be reflected in the model."* RA8

That helped the computer science undergraduate students to understand how data affects the decisions of the intelligent systems we built to address some socio-technical issues.

Discussion

Reflection on Unsafe Conversations Caused Some Discomfort but Increased Self-awareness and Facilitated Desired Behavior Change

Overall, the majority of participants were glad to contribute to the research, although some of them expressed concerns about their privacy. Regarding privacy, our results provided some of the best privacy practices that participants expect to have. Participants mentioned de-identifying their data or summarizing the conversations. It is good that researchers consider not quoting directly and consider Bruckman's [57] levels of disguise when using quotes in their publications. As a consequence of participating in our study, some participants managed the emotional distress of reliving the sensitive interactions by learning from them. In addition, we found that annotators were surprised to see the frequency of risks for teens, some felt uncomfortable but it did not cause them long-term distress. We mentioned in the consent/assent forms that risks to participants are minimal and do not exceed the risks associated with activities found in daily life. However, we mentioned that the online survey includes questions about some sensitive topics in which some people may become anxious or upset when reflecting on their behaviors, well-being, or views. We provided contact information for them to contact counseling and provided help resources available at all times during the study. Although we disclosed to participants that potential risks include feeling a level of discomfort, we did not have the potential risks written for researchers. When conducting sensitive research this question could be asked: "From the human-subjects perspective, could we consider having a consent form for researchers?" One of the useful protecting measures we could have performed to protect researchers before they accept to work on a project is to have a consent form for researchers, explaining the potential benefits and harms in annotating data and conducting the research. That way they would have a better idea and expectations. Trauma-informed frameworks [285, 110, 236] should be integrated with online safety research. This could

be consisted of teaching self-care strategies to overcome emotional fatigue for researchers and empowering participants.

We found that participants' reflections increased their privacy awareness and changed their social media habits. Also, annotators learned more about online risks, reflected on their own past experiences, and gave advice to their loved ones. In the mental health space, researchers [248, 204] have created interventions based on self-reflections and investigated ways that self-reflections help people change their behaviors. For instance, mood-tracking applications bring more self-awareness to people and provide an understanding of previous events and emotions associated with them. This research shows that in addition to mental health and well-being, self-reflections could be helpful for online risk behavior. This demonstrates a potential liability of applying that for online safety space.

Needed Direct Support from Research Team

In addition to the risky research guidelines [29], we also need to consider ways to engage with the research team more frequently. Researchers should not be expected to do independent Inter-rater Reliability Calculations or try to make everything online. Having more in-person support is helpful, even if it is just helping participants in a Zoom call when they are uploading their data. Overall, it is important to make the research more of a hands-on process. As researchers are moving more away from conducting user studies to conducting big data analysis, interacting with people goes out of the picture. Therefore, researchers might not be getting important details such as their private message conversations and moving toward scraping public data.

The data that the researchers obtain in online safety area might be stories of intense suffering, social injustices, or other things that would shock the researcher. It is important to provide an environment for data annotators to mitigate any negative impact on them. We found that providing

help and feedback from the research mentors is very important to support the annotators. Each sensitive research is unique, and challenges and emotional distress related to it might be different in nature [283]. So before starting a research process, it is important to have a well-being care plan to prepare researchers to express any fatigue or trauma that may be experienced and ways to overcome it. There might be some stigma related to expressing emotions, especially in highly technical fields such as Computer Science. Taboo lingers in the academic arenas of Computer Science to maintain professionalism in our research and exposing emotions is against professionalism. So it becomes even more important to provide a safe environment for researchers to express their personal experiences and remote the emotional distance in the research group. Therefore we encourage HCI researchers to utilize strategies from social science researchers and other fields that work more with human-subject. These strategies for researcher self-care include personal self-assessment of emotional risk factors, emotional proximity, and distance, physical health and wellbeing, mental time-out, social support, and enabling environment [283].

Our findings shed light on the importance of contextual cues to both participants and research assistants to flag the data. As a research team, we understood the importance of being human-centered. But we are also getting empirical evidence from participants themselves about the fact that context really mattered; such as relationship type between people involved. For instance, understanding the intention of people is a difficult task, computational researchers tried to detect sarcasm. There is a need to go beyond typical textual embedding and linguistic cues to understanding these types of contextual cues. Also, it is important to not only consider the perspective of participants on getting ground truth but also teaching computer science students to think outside the box; to care about human-centeredness in their future systems development and research. It is not only on the research community to emphasize the importance of human-centeredness in computing, but also it is important to embed that in our educational system as well as participants themselves

Limitations and Future Research

We call researchers to perform ethical research on how to manage the impact on the practitioner and researchers in the challenging context of online safety. The formal best practice guidelines for conducting research on sensitive topics should be established in the HCI community. In this work, we interviewed participants which this approach is prone to self-reported data biases such as social desirability response bias or recall bias [274]. Longitudinal studies could be used to understand youth behavior better, for instance, if they mentioned that their social media behavior changed as the result of flagging their risky data, diary studies could be used to study their long-term behavioral changes with less recall bias. Future research could use facial recognition or sensors to measure the stress levels when participants to through reviewing their uncomfortable/unsafe social media interactions.

Conclusion

Our work provided insights into the perspectives of the participants participating in the research by sharing sensitive social media data and how it impacted them to review their past online risky interactions. Also, our findings shed light on how annotating unsafe conversations from youth had impacts on data annotators and the mental health of the research assistants.

CHAPTER 7: OUTCOMES

In this chapter, we provide a summary of the overall findings, contributions and outcomes of this dissertation. Next, we discuss future directions and end with a conclusion.

Research Summary

In this dissertation study, we focused on providing insights on challenging and risky online sexual interactions in which youth are dealing with, and created an ecologically valid dataset based on adolescents' perspectives on those types of risk and built classifiers to detect sexual risks. Therefore, we answered the following research questions in Chapters 2-5:

- **RQ1-Literature Review:** *What are the trends, gaps, and opportunities in the current literature of computational approaches for online sexual risk detection? How address the gaps within the existing literature and provide recommendations for a research agenda that would advance beyond the state-of-the-art within this research domain?*

  In Chapter 2, we conducted a human-centered systematic literature review to analyze 73 peer-reviewed articles on computational approaches utilizing text or meta-data/multimedia for online sexual risk detection. We identified sexual grooming (75%) is the dominant type of sexual risk detection present in the literature. Furthermore, we found that the majority (93%) of this work has focused on identifying sexual predators after-the-fact, rather than taking more nuanced approaches to identify potential victims and problematic patterns that could be used to prevent victimization before it occurs. Many studies rely on public datasets (82%) and third-party annotators (33%) to establish ground truth and train their algorithms. That showed lack of ecological validity of the available datasets which are not representative

of the users, particularly youth, who are most vulnerable to these risks.

- **RQ2- Study 1:** *What are the key characteristics of online sexual experiences that adolescents seek support for?*

  In Chapter 3, we addressed RQ2 by conducting a thematic content analysis of 4,180 posts by adolescents on an online peer support mental health forum. We found youth used the platform to seek support (83%), connect with others (15%), and give advice (5%) about sexting, their sexual orientation, sexual abuse, and explicit content. Majority of youth received overwhelming amount of unwanted sexual solicitation from strangers while seeking support—to the point that adolescents gave advice to one another on which users to stay away from. Meanwhile, they struggled with how to turn down sexting requests from people they knew. This study shed a light on the importance of the relationship of people involved in sexual risks and contextual matters which informed the next studies.

- **RQ3- Study 2:** *How could we gather ecologically valid and human-centered datasets for training machine learning models?*

  In Chapter 4, we addressed RQ3 by developing a data donation platform for youth to gather ecologically valid datasets based on teens' real world social media data and contextualize their perspective on risk for labeling this dataset. We created a training datasets for adolescent online risk detection employing human-centered design principles.

- **RQ4- Study 3:** *Can we use human-centered machine learning to accurately detect these sexual risk experiences?*

  In Chapter 6, we answered RQ4 by conducting experiments on the dataset gathered in Chapter 4. We built traditional machine learning algorithms and a deep learning algorithm for online sexual risk detection in youth private conversations. For identifying sexual risks at using the whole conversation as the unit of analysis, Convolutional Neural Network (CNN) and Random Forest models outperformed (AUC=0.88). Our experiments showed that classifiers

trained on entire conversations performed better than message-level classifiers (AUC=0.85). We also trained classifiers to detect the severity risk level (i.e., safe, low, medium-high) of a given message with CNN outperforming other models (AUC=0.88). We performed feature analysis and found that contextual features (e.g., age, gender, and relationship type) and Linguistic Inquiry and Word Count (LIWC) contributed the most for accurately detecting sexual conversations that made youth feel uncomfortable or unsafe which provides insights into the important factors and contextual features that advance automated detection of sexual risks within youths' private conversations.

- **RQ5- Study 4:** *Retrospectively, can we ensure the safety and well-being of our research participants and research team when studying online sexual risks by understanding the impact of flagging unsafe content on them?*

  In Chapter 5, we answer RQ4 by conducting 29 interviews with youth (ages 13-21) who donated and annotated their Instagram data for online risks ground truth. Additionally, we interviewed 12 research assistants who annotated the participants private conversations for online risks to investigate how it affected their mental health. We found that youth were comfortable sharing their data to advance the adolescents online safety as long as their data would not be used for making profit or not be shared publicly. Also reviewing their conversations and flagging them for risks taught them valuable lessons and increased their risk awareness. We discovered that research assistants were surprised to observe the frequency and variety of the risks that youth are exposed to, but it did not cause them emotional distress. They learned more about the online risks and privacy issues and how they could keep themselves and their loved ones safe online. Our research provides insights and implications for design for data annotation for online risks and mental health of annotators.

Research Contributions

This dissertation makes several contributions according to Wobbrock's work on research contributions to the fields of Human-Computer Interaction (HCI), adolescent online safety, Human-centered Machine Learning (HCML), Machine Learning (ML), and online peer support [300] as following:

- Our qualitative research makes empirical contributions. First, this contribution was made by providing Human-Centered Lens for Computational Risk Detection Systematic Literature Reviews framework for systematically reviewing computation risk detection literature using a human-centered lens. We provided in-depth synthesis of the current state-of-the-art, trends, and gaps in computational approaches for online sexual risk detection and recommendations for a research agenda that would advance beyond the state-of-the-art within this research domain. Second, in Study 1, we provided a deeper understanding of peer support seeking behaviors for adolescents online sexual experiences and uncovers the challenges of online peer support and makes recommendations for design. Third, in Study 4, we provided insights on the ethical treatment of research participants who engage in more than minimal risk research and research assistants that annotated unsafe sensitive online interactions.

- Our study 2 and 3 make dataset and artifact contributions by utilizing human-centered approach to build ecologically valid dataset based on digital trace data from youth and their perspective of sexual risks. Lastly, automated machine learning models were developed to predict the presence and severity of sexual risks in conversations of youth which could act as a key element in ensuring the online safety of youth and young adults.

Future Research Directions

For each study, the future research directions are stated in its corresponding chapter. Researchers across disciplines should work together to utilize human knowledge and user centred design to create solutions for online sexual risks for youth. As mentioned in previous chapters, more multi-modal approaches for risk detection should be developed to consider more context. The detection models built in this study should be tested and evaluated in real-world applications. In addition, risk mitigation applications should be designed to make the right actions as soon as a risk being detected. The ethics, fairness, and accountability of those systems should be studied and discussed in future endeavors. In study 3, the trade-off between using traditional machine learning models and deep learning models has been discussed. Future studies should work on the explainability of the deep learning models especially for sensitive applications such as online risk detection to reveal how algorithms learn.

Conclusion

Overall, this work contributed to the dire societal issues of adolescent online safety by detecting and providing insights relevant to youths' risky online sexual interactions. More broadly, this dissertation makes contributions to the fields of Human-Computer Interaction, Machine Learning, and adolescent online safety. It does this by synthesizing research on existing computational approaches for online sexual risk detection to set a research agenda for the field. Additionally, it provides a deeper understanding of peer support-seeking behaviors for adolescents' online sexual experiences, which uncovers the challenges of online peer support and makes recommendations for design. By utilizing human-centered principles, we collected an ecologically valid dataset of digital trace data from youth and leveraged this dataset to build machine learning models to predict

the presence and severity of sexual risks in conversations of youth.

APPENDIX A: LITERATURE REVIEW ADDITIONAL TABLES

Table A.1: Media Type

Media Types	**Counts (Percent)**	**References**
Text	47 (64%)	[146, 92, 177, 192, 250, 305, 51, 237, 175, 244, 114, 213, 135, 86, 225, 47, 99, 27, 61, 279, 138, 91, 109, 74, 145, 68, 260, 144, 243, 45, 108, 190, 282, 124, 189, 302, 24, 251, 184, 262, 188, 218, 238, 85, 49, 309, 217]
Text and Meta-data	16 (21%)	[219, 174, 45, 15, 214, 164, 224, 67**?** , 130, 163, 136, 89, 171, 155, 120]
Text and Meta-data and image	7 (10%)	[288, 254, 132, 133, 131, 272, 168]
Meta-data	2 (2%)	[271]
Text and image	1 (1%)	[**?**]

Table A.2: Data Types

Datasets	**Counts (Percent)**	**Publications**
Perverted Justice	22 (30%)	[218, 155, 188, 184, 189, 89, 190, 45, 163, 138, 61, 27, 225, 224, 114, 237, 51, 305, 250, 192, 174, 238]
PAN-2012	16 (22%)	[146, 92, 192, 49, 51, 175, 309, 85, 86, 27, 279, 91, 109, 217, 282, 124]
Combined Chat datasets	10 (14%)	[46, 114, 214, 135, 225, 47, 45, 24, 155, 309]
Social Media	8 (11%)	[**?** 214, 260, 108, 145, 262**?** , 120]
Advertisements	8 (11%)	[288, 254, 132, 133, 131, 130, 168, 251]
Queries	4 (5%)	[244, 213, 164, 99]
Private Chat data	3 (4%)	[271, 74, 136]
SafeCity	3 (4%)	[144, 302, 171]
Games	2 (3%)	[219, 67]
Forums	2 (3%)	[214, 158]
Anonymous Platforms	2 (3%)	[15, 68]
Blogs	2 (3%)	[214, 27]
Generated by Participants	1 (1%)	[243]

Table A.3: Ground Truth Annotators

Annotators	**Counts (Percent)**	**Publications**
Existing	32 (44%)	[146, 92, 192, 250, 305, 51, 175, 114, 85, 86, 224, 47, 27, 279**?** , 91, 272, 217, 168, 144, 45, 190, 282, 124, 89, 24, 188, 218, 49, 214, 309, 138**?**]
Outsiders	24 (33%)	Researchers (26%): [237, 15, 225, 254, 132, 61, 133, 109, 74, 145, 68, 108, 184, 171, 262, 155, 120, 238, 250] Moderators (3%): [67, 131] Clinical (1%): [108] Crowd-source (1%): [243]
Auto	17 (23%)	[146, 219, 174, 46, 288, 244, 164, 132, 99, 130, 163, 260, 108, 136, 189, 302, 251]
Insiders	4 (5%)	[145, 158, 260, 108]

Table A.4: Features

Features	**Counts (Percent)**	**Publications**
Textual	62 (85%)	[146, 92, 219, 192, 250, 49, 51, 237, 175, 15, 288, 309, 114, 214, 213, 85, 135, 86, 224, 225, 47, 254, 27, 61, 279**?** , 67**?** , 138, 131, 133, 163, 74, 272, 217, 145, 168, 68, 158, 144, 243, 45, 108, 190, 282, 124, 89, 189, 302, 24, 251, 184, 171, 188, 155, 120, 218, 238, 305, 46, 109]
User	20 (27%)	[92, 288, 47, 132, 27, 271, 279, 67, 138, 133, 130, 91, 109, 74, 217, 168, 158, 124, 184, 262]
Time/Location	14 (19%)	[177, 288, 214, 254, 132, 271, 279, 133, 131, 130, 168, 89, 184, 171]
Semantic	12 (16%)	[177, 174, 46, 175, 15, 214, 213, 86, 61**?** , 89, 189]
Style	11 (15%)	[174, 305, 237, 15, 61, 67, 138, 68, 260, 45, 188]
Behavioral	10 (14%)	[237, 114, 271, 279, 67, 217, 45, 124, 184, 155]
Keyword Extraction	10 (14%)	[219, 244, 214, 135, 99**?** , 272, 136, 171, 155]
Syntactic	8 (11%)	[49, 15, 214, 85, 27, 61**?** , 272]
Sentiment	7 (10%)	[174, 47, 61, 67, 272, 260, 45]
Images	5 (7%)	[288, 254**?** , 133, 272]
Network	4 (5%)	[164, 271, 74, 158]
Topic Modeling	3 (4%)	[305, 175, 15, 251]
Relationships	3 (4%)	[164, 45, 184]

Table A.5: Approaches

Approaches	**Counts (Percent)**	**Publications**
Traditional ML	48 (66%)	Supervised (53%): [92, 250, 49, 305, 51, 237, 175, 15, 288, 309, 114, 214, 213, 47, 27, 61, 279**?** , 67, 138, 163, 91**?** , 74, 272, 217, 260, 243, 45, 190, 282, 124, 189, 24, 184, 262, 188, 155, 120, 218] Unsupervised (6%): [174, 254, 271, 136, 192, 224, 217, 184] Semi-supervised (3%): [86, 168, 158]
Hand-crafted or Rule-based	13 (18%)	[219, 244, 135, 164, 225, 279**?** , 130, 68, 89, 184, 109, 238]
Deep Learning	11 (15%)	[146, 192, 85, 145, 144, 108, 302, 251, 171, 238, 305]
Graph or Network-based	5 (7%)	[132, 133, 131, 99, 168]
System Architecture	5 (7%)	[177, 219, 288, 164, 254]

Table A.6: Algorithms Names

Algorithms	**Counts (Percent)**	**Publications**
Support Vector Machine (SVM)	27 (37%)	[92, 250, 49, 51, 237, 175, 15, 288, 114, 214, 213, 86, 224, 47, 61, 279**?** , 67, 74, 217, 158, 260, 243, 282, 189, 120, 218]
Naive Bayes (including Multinomial, Gaussian, and Bernoulli) (NB)	16 (22%)	[92, 49, 51, 15, 309, 86, 279, 67, 74, 243, 45, 190, 189, 124, 24, 188]
Neural Networks (NN) (including CNN, RNN, LSTM)	15 (21%)	[146, 92, 192, 51, 114, 85, 67, 145, 144, 108, 282, 302, 171, 262, 218, 238, 305]
Regressions (including Logistic (LR), Ridge (RR), Bayesian (BR))	10 (14%)	[92, 51, 309, 213, 225, 67, 24, 51, 15, 138]
K Nearest Neighbor (KNN)	9 (12%)	[92, 114, 254, 67, 158, 260, 184, 218, 189]
Decision Tree (DT)	8 (11%)	[92, 279, 67, 243, 184, 188, 155, 49]
Clustering algorithms (including Agglomerative, AC, KMEANS, AGG, BHC, GMM)	7 (10%)	[224, 61, 168, 158, 136, 155, 271]
Language Model (including BERT)	4 (5%)	[174, 46, 163, 251]
Random Forest (RF)	4 (5%)	[92, 309, 120, 49]
AdaBoost (AB)	3 (4%)	[309, 27, 24]
Contrastive Pessimistic Likelihood Estimation (CPLE)	2 (3%)	[61, 158]
Self-training (ST)	2 (3%)	[272, 260]
Multiple Sequence Alignment (MSA)	1 (1%)	[224]
Local Interpretable Model-Agnostic Explanation (LIME)	1 (1%)	[144]
Linear classifier (LINEAR)	1 (1%)	[305]
RIPPER rule-learning algorithm (RIPPER)	1 (1%)	[184]
Ring Based Classifier (RING)	1 (1%)	[91]
Temporal Relational Semantic Systems (TRSS)	1 (1%)	[89]
Mean Variance	1 (1%)	[109]

Table A.7: Granularity Level

Granularity Level	**Counts (Percent)**	**Publications**
Users	27 (36%)	[146, 92, 219, 192, 51, 237, 175, 214, 85, 135, 224, 47, 27, 271, 279, 67, 138, 109, 217, 68, 45, 282, 124, 184, 155, 218, 86]
Conversations	15 (20%)	[92, 51, 114, 225**?** , 163, 217, 282, 188, 86, 309, 74]
Patterns	9 (12%)	[305, 46, 86, 224, 61, 91, 190, 89, 189, 155, 174, 49]
Lines	8 (11%)	[219, 51, 15, 214, 279**?** , 217, 243, 190]
Levels	2 (3%)	[250, 238]

Table A.8: Output Types

Output Types	**Counts (Percent)**	**Publications**
Binary classification	44 (60%)	[92, 219, 192, 250, 51, 237, 175, 114, 214, 213, 85, 164, 225, 47, 132, 27, 61, 279, 67, 133, 91, 109, 217, 68, 158, 260, 144, 243, 45, 108, 190, 282, 124, 302, 24, 251, 184, 188, 218, 49, 309, 74]
Multi-class classification	24 (33%)	[146, 45, 15, 288, 309, 244, 224, 99**?** , 138, 130, 163, 272, 145, 144, 24, 171, 262, 188, 155, 120, 218, 238, 189]
Clustering	7 (10%)	[135, 254, 271, 133, 168, 136, 155]
Stage detection	5 (7%)	[174, 305, 224, 225, 61]

APPENDIX B: IRB APPROVAL (STUDY 2, 3, and 4)

UNIVERSITY OF CENTRAL FLORIDA

Institutional Review Board
FWA00000351
IRB00001138, IRB00012110
Office of Research
12201 Research Parkway
Orlando, FL 32826-3246

APPROVAL

March 23, 2022

Dear Pamela Wisniewski:

On 11/22/2019, the IRB reviewed the following submission:

Type of Review:	Initial Study
Title:	Social Media and Youth Study
Principal Investigator:	Pamela Wisniewski
Co-Investigator	Afsaneh Razi
IRB ID:	STUDY00001136
Funding:	Name: Natl Science Fdn (NSF), Grant Office ID: 1065217, Funding Source ID: 1827700
Grant ID:	1065217;
IND, IDE, or HDE:	None
Documents Reviewed:	• Flyer , Category: Recruitment Materials; • Flyer 2, Category: Recruitment Materials; • IGDD _FAQ_About Us_pages_FINAL.docx, Category: Other; • IGDD end of survey email - Eligible.rtf, Category: Other; • IGDD end of survey email - Not Eligible.rtf, Category: Other; • IGDD reminder email_FINAL.rtf, Category: Other; • IGDD reminder email_Frequency_FINAL.rtf, Category: Other; • IGDD_Adult_Consent, Category: Consent Form; • IGDD_Data_Survey_FINAL.pptx, Category: Survey / Questionnaire; • IGDD_Help_Resources.docx, Category: Other; • IGDD_Parental_Consent, Category: Consent Form; • IGDD_Pre-Screening_FINAL.docx, Category: Survey / Questionnaire; • IGDD_Recruitment_Email- Organizations_FINAL.docx, Category: Recruitment Materials; • IGDD_Recruitment_Email_FINAL.docx, Category: Recruitment Materials; • IGDD_Recruitment_Flyer_Final.pptx, Category: Recruitment Materials; • IGDD_Recruitment_Flyer_Social Media.pptx, Category: Recruitment Materials; • IGDD_Teen_Assent, Category: Consent Form; • Instructions_Part1_Android.pdf, Category: Debriefing Form; • Instructions_Part1_iOS.pdf, Category: Debriefing Form; • Instructions_Part1_Web.pdf, Category: Other; • Instructions_Part2_Web.pdf, Category: Other; • NSF PFI Phases 1+2 - IRB Protocol, Category: IRB Protocol;

	• ParentingYourTeen_Handout1.pdf, Category: Other; • Phase 2 Adult Consent.pdf, Category: Consent Form; • Phase 2 Parent Consent.pdf, Category: Consent Form; • Phase 2 Session Script , Category: Interview / Focus Questions; • Phase 2 Teen Assent.pdf, Category: Consent Form; • Phase_1__Re-Consent.pdf, Category: Consent Form; • Phase_2_Transition_Screen(1).pptx, Category: Recruitment Materials; • Phase_2_v2_Scheduling_Email - Copy.docx, Category: Recruitment Materials; • Reconsent-phase2.docx, Category: Other; • Social_Media_and_Youth_Survey.docx, Category: Survey / Questionnaire; • Youth Services and Resources for Youth and Teens _ National Runaway Safeline.pdf, Category: Other

The IRB approved the protocol from 11/22/2019 to 3/8/2023.

In conducting this protocol, you are required to follow the requirements listed in the Investigator Manual (HRP-103), which can be found by navigating to the IRB Library within the IRB system. Guidance on submitting Modifications and a Continuing Review or Administrative Check-in are detailed in the manual. When you have completed your research, please submit a Study Closure request so that IRB records will be accurate.

If you have any questions, please contact the UCF IRB at 407-823-2901 or irb@ucf.edu. Please include your project title and IRB number in all correspondence with this office.

Sincerely,

Katie Kilgore
Designated Reviewer

APPENDIX C: INFORMED CONSENTS AND ASSENT (STUDY 2, 3, and 4)

UNIVERSITY OF CENTRAL FLORIDA

Title of research study: Social Media and Youth Study

Investigator: Pamela Wisniewski, Ph.D.
Co-investigators: Munmun De Choudhury, Ph.D., Gianluca Stringhini, Ph.D., Elizabeth Cauffman, Ph.D., Kimberly Gryglewicz, Ph.D., Afsaneh Razi, Heidi Hartikainen Ph.D., Zainab Agha, Neeraj Chatlani, Seunghyun Kim

How to Return this Consent Form:
Your (as the parent or legal guardian of a minor) electronic signature is required to enroll in this study. Once you submit the form with your electronic signature, this will indicate your agreement to for your teen to participate in the research as described in more detail below. If you would like, you can download or print this consent form for your records.

Key Information:
The following is a short summary of this study to help you decide whether or not to allow your teen to be a part of this study. More detailed information is listed later on in this form.

Why is my teen being invited to take part in a research study?
Your teen is being invited take part in a research study because he or she is between the ages of 13-17 years old, has an active Instagram account, has received at least 2 direct message conversations with someone that made your teen or someone else feel uncomfortable or unsafe, and is willing to share their Instagram data for the purpose of research. Active is defined as having an Instagram account for at least 3 months and exchanging direct messages with at least 15 people. To be eligible, your teen must currently reside in the United States and speak fluent English. An uncomfortable or unsafe interaction may include:

- **Nudity/porn**: Photos or videos of a nude or partially nude person.
- **Sexual messages or Solicitations**: Someone sent me a sexual message ("Sexting"). Someone asked me to send them a sexual message, revealing, or naked photo of myself.
- **Harassment**: Messages that contain credible threats, targets to degrade or shame someone, contains personal information to blackmail or harass someone, threats 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 or photos or videos of extreme violence, or encourage violence or attacks anyone based on their religious, ethnic or sexual background
- **Sale or promotion of illegal activities:** Messages contain promoting the use, sell, or distributing illegal material such as drugs.
- **Self-injury:** Messages encouraging or promoting self-injury, which includes suicide, cutting and eating disorders.
- **Other:** Other situations that could potentially lead to emotional or physical harm.

Why is this research being done?
As teens engage with others on social media, sometimes they encounter situations that make them feel uncomfortable or unsafe. We want to better understand these types of experiences teens encounter on social media (specifically on Instagram), so that we can design interventions that help teens feel safer and empowered online.

How long will the research last and what will my teen need to do?
If you sign this consent form, your teen will be enrolled in the study . They will be asked to download their Instagram data, take a web-based survey, upload their Instagram data, and take a survey based on their Instagram experiences. The entire process should take from 1 hour to 3 days to complete, depending on how long it takes Instagram to prepare their data file for download. Your teen will receive email reminders to complete each step of this process should they need to leave the study and come back later. You may also be contacted after your teen completes this study, to participate in a follow-up study alongside your teen.

More detailed information about the study procedures can be found under **"What happens if I say yes, I want my teen to be in this research?"**

Is there any way being in this study could be bad for my teen?
The risks of participation are minimal and do not exceed the risks associated with activities found in daily life. However, the online survey includes questions about some sensitive topics related to social media use, such as cyberbullying, unwanted sexual experiences, mental health including depression and self-harm, and risky behaviors. Some people may become anxious or upset when answering questions about their behaviors, well-being or views. If you believe your teen need counseling, please contact 1-877-SAMHSA7 (1-877-726-4727), or refer to the resources provided in the "Help Resources" section available at all times during the study.

Will being in this study help my teen in any way?
Possible benefits include that participating in this study may increase your teen's awareness of their social media activities and start open conversations about experiences they are having online.

What happens if I do not want my teen to be in this research?
Participation in research is completely voluntary. You can decide to have your teen participate or not to participate.

Detailed Information: The following is more detailed information about this study in addition to the information listed above.

What should I know about a research study?

- Someone could explain this research study to you and your teen.
- Whether or not you allow your teen to take part is up to you.
- You can choose not to allow your teen to take part.
- You can agree to allow your teen to take part and later change your mind.

- Your decision will not be held against you or your teen.
- You can ask all the questions you want before you decide.

Who can I talk to?
If you have questions, concerns, or complaints, or think the research has hurt your teen, talk to the research team by contacting Dr. Pamela Wisniewski at pamwis@ucf.edu.
This research has been reviewed and approved by an Institutional Review Board ("IRB"). You may talk to them at 407-823-2901or irb@ucf.edu if:

- Your questions, concerns, or complaints are not being answered by the research team.
- You cannot reach the research team.
- You want to talk to someone besides the research team.
- You have questions about your teen's rights as a research subject.
- You want to get information or provide input about this research.

How many people will be studied?
For Part 1, we expect to enroll up to 1500 participants. Out of these, up to 75 teens or young adults will be recruited to participate in Part 2 of the study.

What happens if I say yes, I want my teen to be in this research?

- After agreeing to participate in this study, your teen will be given instructions on how to download their Instagram data file. This process will take no more than 5 minutes to complete. However, it may take Instagram up to 48 hours to fulfill the request. Your teen will later be asked to upload this file to our system. This file includes their Instagram posts, direct messages, and photos. Posts made by their friends are not included in this file.

- Next, they will be asked to complete a web-based survey about their social media usage and personal experiences which includes questions about social media use, cyberbullying, unwanted sexual experiences, mental health including depression and self-harm, and risky behaviors. This survey will take approximately 30-60 minutes to complete.

- Your teen will receive an email from Instagram to download their data file within 48 hours of their initial request. Once downloaded, your teen will be asked to upload this file to our secure system. This process should take approximately 5-30 minutes to complete depending on the file size. **Please do not upload any Instagram file that may contain any visual depiction of sexually explicit conduct involving a minor (persons less than 18 years old). Federal law requires us to report child pornography that we find in the data to the proper authorities.**

- Once the Instagram data has been uploaded, your teen will be asked to identify direct messages that made them feel uncomfortable or unsafe in the past. We will ask your teen to answer questions to help us better understand these experiences. This survey should take approximately 30-90 minutes to complete.

- In case your teen closes the browser he/she will receive an email with a link to continue and will receive reminder emails every 48 hours to continue. Participants can

unsubscribe from emails at any time by hitting unsubscribe link in the emails. Unsubscribing from emails doesn't affect participation in the study.

- We ask that you thoughtfully consider your teen's privacy and discuss whether or not your teen would feel more comfortable if they were allowed to answer the survey questions without your direct supervision. Our intent is to allow teens to be open and honest in their responses and not feel worried that they may get in trouble by participating in this research study.

- Your teen may receive an offer to participate in phase 2 of the study which includes 30 minutes interview session that includes questions about how they felt participating in phase 1 of the study. If they receive this offer, they may choose to opt into the study, in which case the researchers will contact them in order to begin phase 2. If they choose not to participate, they may opt out, in which case they will not be contacted.

What happens if I say yes, but I change my mind later?
You or your teen can choose to leave the study at any time, and it will not be held against you or your teen. If you or your teen decides to leave the study prior to completion, your teen will not receive any compensation. If you or your teen would like to completely withdraw from the study and not have the data collected used for the purpose of this research, please email Dr. Pamela Wisniewski at pamwis@ucf.edu to make this request within 48 hours of the study completion. If your teen partially completes the study but does not withdraw, we will use the data they have already shared.

What happens to the information collected for the research?

The information collected for this study will be kept confidential to the extent permitted by law. We take your teen's privacy seriously and will treat their data with the utmost care. The Instagram data collected for this study may include personally identifiable information (e.g., names, phone numbers, addresses, etc.). However, no results will be published from this study that would allow your teen to be personally identified by the information disclosed. We will follow best practices to ensure that all data is encrypted and stored securely. Efforts will be made to limit the use and disclosure of your teen's personal information included in this research study to people who have a need to review this information. However, we **cannot** promise complete secrecy. Organizations that may inspect and copy the participant's information include the IRB and other representatives of this organization to ensure the ethical compliance of this research procedure is being upheld. As the research sponsor, the National Science Foundation may also inspect the data.

If data collected from this study strongly indicates that your teen is at serious risk of physical injury, sexual abuse, mental injury, or physical neglect we, as mandated reporters, are required to report these types of imminent risks to the proper authorities. However, the data will not be screened directly for such risks and will only be reported in the case that an imminent risk is found. The content of the Instagram data file may include information shared with your teen by other users. This data will be treated as

part of your teen's data and will be under the same requirements for mandated reporting as your teen's own data.

All data collected must be retained for a minimum of five years per Florida statute. De-identified data must be retained for a minimum of ten years per the National Science Foundation's requirement for data retention for funded research. **These data retention laws apply even in the case that you or your teen choose to withdraw from the study.**

The data collected for this study will not be shared publicly but may be shared with a limited group of researchers who are working closely with the co-investigators. Those researchers will be required to comply with all assurances and regulations set forth by UCF's IRB. Portions of the social media data may also be shared with researchers requesting third-party access. The researchers will remove Instagram usernames and participants' contact information (including participants' names and email addresses) prior to sharing the data. To gain third-party access, researchers must show an established record of relevant, published research to validate why they should have access to this data, receive IRB approval from their home institutions, and sign a binding data confidentiality agreement with the primary investigator. Under no circumstances will third-parties be allowed to re-distribute any data collected from this study.

If you want your teen's information to only be used for this research study, please request that by emailing Dr. Pamela Wisniewski at pamwis@ucf.edu within 2 weeks after your teen completed the study.

What else do I need to know?

If your teen completes this research study data upload and survey and their data is verified to match the study inclusion criteria, we will pay them a $50 Amazon gift. Gift cards will be sent via email to the address you specify after agreeing to this consent form. Your teen needs to complete the study and provide her/his Instagram data in other to be eligible to receive the gift card. If she/he exits the study after completing the survey but does not provide her/his Instagram data, she/he will not receive the gift card. Gift cards will be sent within 60 days of completing the study once the data has been verified based on the pre-screening eligibility requirements of this study. If your teen's data is not found to match the study inclusion criteria, you will be contacted to be told why their data did not meet inclusion criteria.

Additionally, you may be contacted to participate in a follow-up study involving both you and your teen, depending on your eligibility.

This research is covered by a Certificate of Confidentiality from the National Institutes of Health. This means that the researchers cannot release or use information, documents, or samples that may identify you in any action or suit unless you say it is okay. They also cannot provide them as evidence unless you have agreed. This protection includes federal, state, or local civil, criminal, administrative, legislative, or other proceedings. An example would be a court subpoena.

There are some important things that you need to know. The Certificate DOES NOT stop reporting that federal, state or local laws require. Some examples are laws that require reporting of child or elder abuse, child pornography, some communicable diseases, and threats to harm yourself or others. The Certificate CANNOT BE USED to stop a sponsoring United States federal or state government agency from checking records or evaluating programs. The

Certificate also DOES NOT prevent your information from being used for other research if allowed by federal regulations.

Researchers may release information about you when you say it is okay. For example, you may give them permission to release information to insurers, medical providers or any other persons not connected with the research. The Certificate of Confidentiality does not stop you from willingly releasing information about your involvement in this research. It also does not prevent you from having access to your own information.

This research is being funded by the National Science Foundation. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the U.S. National Science Foundation. Thank you for allowing your teen to be part of our research study!

Signature Block for Children

Your signature documents your permission for the named child to take part in this research.

Printed name of child

Signature of parent Date

Printed name of parent

Signature of parent Date

Printed name of parent

If signature of second parent not obtained, indicate why: (select one):

- ❑ Second parent is deceased
- ❑ Second parent is unknown
- ❑ Second parent is incompetent
- ❑ Second parent is not reasonably available
- ❑ Only one parent has legal responsibility for the care and custody of the child

Email Address

Please provide your email address (parent's email address not the teen's email address) for our record:

_____

Please print this form for your records.

UNIVERSITY OF CENTRAL FLORIDA

Title of research study: Social Media and Youth Study

Investigator: Pamela Wisniewski, Ph.D.
Co-investigators: Munmun De Choudhury, Ph.D., Gianluca Stringhini, Ph.D., Elizabeth Cauffman, Ph.D., Kimberly Gryglewicz, Ph.D. Afsaneh Razi, Heidi Hartikainen Ph.D., Zainab Agha, Neeraj Chatlani, Seunghyun Kim

How to Return this Consent Form:
You need to electronically sign this form to be in this study. Once you submit the form with your electronic signature, this will mean that you agree to participate in the research that is described below.

Key Information:
The following is a short summary of this study to help you decide if you want to be a part of this study. More detailed information is listed later on in this form.

Why am I being invited to take part in a research study?
You are being invited take part in a research study because you are between the ages of 13-17 years old, have an active Instagram account, and you participated in at least 2 direct message conversations with someone that made you or someone else feel uncomfortable or unsafe. You also said that you are willing to share your Instagram data for the purpose of research. 'Active' means that you have had an Instagram account for at least 3 months and have exchanged direct messages with at least 15 people. To be eligible, you must currently reside in the United States and speak fluent English. An uncomfortable or unsafe interaction may include:

- **Nudity/porn**: Photos or videos of a nude or partially nude person.
- **Sexual messages or Solicitations**: Someone sent me a sexual message ("Sexting"). Someone asked me to send them a sexual message, revealing, or naked photo of myself.
- **Harassment**: Messages that contain credible threats, targets to degrade or shame someone, contains personal information to blackmail or harass someone, threats 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 or photos or videos of extreme violence, or encourage violence or attacks anyone based on their religious, ethnic or sexual background
- **Sale or promotion of illegal activities:** Messages contain promoting the use, sell, or distributing illegal material such as drugs.
- **Self-injury:** Messages encouraging or promoting self-injury, which includes suicide, cutting and eating disorders.
- **Other:** Other situations that could potentially lead to emotional or physical harm.

Why is this research being done?
As teens engage with others on social media, sometimes they have experiences that make them feel uncomfortable or unsafe. We want to better understand the types of experiences that teens have on social media (specifically on Instagram), so that we can design tools that help teens feel safer and empowered online.

How long will the research last and what will my teen need to do?
If you sign this consent form, you will be enrolled in the study. You will be asked to download your Instagram data, take a web-based survey, upload your Instagram data, and take another survey based on your Instagram experiences. The entire process should take from 3 hours to 3 days to complete, depending on how long it takes Instagram to send your data file. You will receive email reminders to complete each step of the study, if you need to leave the study and come back later. You may also be contacted after you complete this study, to be in another study with your parent, if you want to.

More detailed information about the study can be found under **"What happens if I say yes, I want to be in this research?"**

Is there any way being in this study could be bad for me?
The risks of participation are minimal and do not exceed risks found in daily life activities. . However, the online survey includes questions about some sensitive topics related to social media use, such as cyberbullying, unwanted sexual experiences, mental health including depression and self-harm, and risky behaviors. Some people may become anxious or upset when answering questions about their personal experiences. If you believe you need counseling, talk to your parent, or contact 1-877-SAMHSA7 (1-877-726-4727) or refer to the resources provided in the "Help Resources" section available at all times during the study.

Will being in this study help me in any way?
Being in this study may increase your awareness of your social media activities and help you have honest conversations about the experiences that you are having online.

What happens if I do not want to be in this study?
Participation in research is completely voluntary. You can decide to participate, or not to participate.

Detailed Information:

The following is more detailed information about this study in addition to the information listed above.

What should I know about a research study?

- Someone should explain this research study to you and your parent.
- Whether or not you decide to take part in this study is up to you.
- You can agree to take part and later change your mind if you decide not to take part.
- Your decision will not be held against you.
- You can ask all the questions you want before you decide.

Who can I talk to?

If you have questions, concerns, or complaints, or think the research has hurt you, your parent can talk to the research team by contacting Dr. Pamela Wisniewski at pamwis@ucf.edu.
This research has been reviewed and approved by an Institutional Review Board ("IRB"). Your parent may talk to them at 407-823-2901or irb@ucf.edu if:

- Your questions, concerns, or complaints are not being answered by the research team.
- You cannot reach the research team.
- You want to talk to someone besides the research team.
- You have questions about your rights as a research subject.
- You want to get information or provide input about this research.

How many people will be studied?

For Part 1, we expect to enroll up to 1500 participants. Out of these, up to 75 teens or young adults will be recruited to participate in Part 2 of the study.

What happens if I say yes, I want to be in this research?

- After agreeing to participate in this study, you will be given instructions on how to download your Instagram data file. This will take no more than 5 minutes to complete. However, it may take Instagram up to 48 hours to create your data file. You will later be asked to upload this file to our system. This file includes your Instagram posts, direct messages, and photos. Posts made by your friends are not included in this file.

- Next, you will be asked to complete a web-based survey about your social media usage and personal experiences. This includes questions about social media use, cyberbullying, unwanted sexual experiences, mental health including depression and self-harm, and risky behaviors. This survey will take approximately 30-60 minutes to complete.

- You will receive an email from Instagram to download their data file within 48 hours of your request. Once downloaded, you will be asked to upload this file to our secure system. This process should take approximately 5-30 minutes to complete depending on the file size. **Please do not upload any Instagram file that may contain any visual depiction of sexually explicit conduct involving a minor (persons less than 18 years old). Federal law requires us to report child pornography that we find in the data to the proper authorities.**

- Once the Instagram data has been uploaded, you will be asked to identify direct messages that made you feel uncomfortable or unsafe in the past. We will also ask you to answer questions to help us better understand these experiences. This survey should take approximately 30-90 minutes to complete.

- In case you close the browser, you will receive an email with a link to continue and will receive reminder emails every 48 hours to continue. You can unsubscribe from these emails at any time by clicking on the unsubscribe link in the emails. Unsubscribing from emails doesn't affect participation in the study.

- We ask that you thoughtfully consider your own privacy and discuss whether or not you would feel more comfortable if you were allowed to answer the survey questions without your parent's direct supervision. Our intent is to allow you to be open and honest in your responses and not feel worried that you may get in trouble by participating in this research study.
- You may receive an offer to participate in phase 2 of the study which includes 30 minutes interview session that includes questions about how you felt participating in phase 1 of the study. If you receive this offer, you may choose to opt into the study, in which case the researchers will contact you in order to begin phase 2. If you choose not to participate, you may opt out, in which case you will not be contacted.

What happens if I say yes, but I change my mind later?
You or your parent can choose to leave the study at any time, and it will not be held against you. If you decide to leave the study before you finish, you will not receive any compensation. If you would like to completely withdraw from the study and not have the data collected used for the research, please ask your parent to email Dr. Pamela Wisniewski at pamwis@ucf.edu to make this request. If you partially complete the study but do not withdraw, we will use the data you have already shared.

What happens to the information collected for the research?
The information collected for this study will be kept confidential to the extent permitted by law. We take your privacy seriously and will treat your data with the utmost care. The Instagram data collected for this study may include personally identifiable information (e.g., names, phone numbers, addresses, etc.). However, no results will be published from this study that would allow you to be personally identified by the information disclosed. We will follow best practices to ensure that all data is encrypted and stored securely. Efforts will be made to limit the use and disclosure of your personal information included in this research study to people who have a need to review this information. However, we **cannot** promise complete secrecy. Organizations that may inspect and copy your information include the IRB and other representatives of this organization to ensure the ethical compliance of this research procedure is being upheld. As the research sponsor, the National Science Foundation may also inspect the data.

If data collected from this study strongly indicates that you are at serious risk of physical injury, sexual abuse, mental injury, or physical neglect we, as mandated reporters, are required to report these types of imminent risks to the proper authorities. However, the data will not be screened directly for such risks and will only be reported in the case that an imminent risk is found. The content of the Instagram data file may include information shared with you by other users. This data will be treated as part of your own data and will be under the same requirements for mandated reporting as your own data.

All data collected must be retained for a minimum of five years per Florida statute. De-identified data must be retained for a minimum of ten years per the National Science Foundation's requirement for data retention for funded research. **These data retention laws**

apply even in the case that you or your parent choose to withdraw from the study.

The data collected for this study will not be shared publicly but may be shared with a limited group of researchers who are working closely with the co-investigators. Those researchers will be required to comply with all assurances and regulations set forth by UCF's IRB. Portions of the social media data may also be shared with researchers requesting third-party access. The researchers will remove Instagram usernames and participants' contact information (including participants' names and email addresses) prior to sharing the data. To gain third-party access, researchers must show an established record of relevant, published research to validate why they should have access to this data, receive IRB approval from their home institutions, and sign a binding data confidentiality agreement with the primary investigator. Under no circumstances will third-parties be allowed to re-distribute any data collected from this study.

If you want your social media data to be used only for this research study, please tell your parent to email Dr. Pamela Wisniewski at pamwis@ucf.edu, within 2 weeks after you complete the study.

What else do I need to know?

If you complete this research study data upload and survey and your data is verified to match the study inclusion criteria, we will pay you a $50 Amazon gift card for your time spent completing this study. Gift cards will be sent to the email address you enter after agreeing to this assent form. You need to complete the study and provide your Instagram data in order to be eligible to receive the gift card. If you exit the study after completing the survey but do not provide your Instagram data, you will not receive the gift card. Gift cards will be sent within 60 days of completing the study once the data has been checked by our researchers. Additionally, you may be contacted to be in another study with your parent, based on the results of this study. If your data is not found to match the study inclusion criteria, you will be contacted to be told why your data did not meet the inclusion criteria.

This research is covered by a Certificate of Confidentiality from the National Institutes of Health. This means that the researchers cannot release or use information, documents, or samples that may identify you in any action or suit unless you and your parents say it is okay. They also cannot provide them as evidence unless you have agreed.

There are some important things that you need to know. The Certificate DOES NOT stop reporting that federal, state or local laws require. Some examples are laws that require reporting of child or elder abuse, child pornography, some communicable diseases, and threats to harm yourself or others. The Certificate CANNOT BE USED to stop a sponsoring United States federal or state government agency from checking records or evaluating programs. The Certificate also DOES NOT prevent your information from being used for other research if allowed by federal regulations.

Researchers may release information about you when your parent says it is okay. For example, your parent may give them permission to release information to insurers, medical providers or any other persons not connected with the research. The Certificate of Confidentiality does not stop you from willingly releasing information about your involvement in this study. It also does not prevent you from having access to your own information.

This research is being funded by the National Science Foundation. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the U.S. National Science Foundation.

After signing this form, choose "Print" from your browser menu or use the "Ctrl+P" keyboard shortcut to print this document. You may also save as a PDF for your records

Thank you for being part of our research study!

Signature Block

Your signature documents your assent to take part in this research.

Printed name of teen

Date

Assent

❑ By checking the boxes below, you (teen) are providing your assent to voluntarily participate in this research study with your parent.

Email Address

(Please provide the email address where you would like reminder notifications and the gift card to be sent. This needs to be teen's email address.)

_____

Please print this form for your records.

UNIVERSITY OF CENTRAL FLORIDA

Title of research study: *Social Media and Youth Study*

Investigator: *Pamela Wisniewski, Ph.D.*

Co-investigators: Munmun De Choudhury, Ph.D., Gianluca Stringhini, Ph.D., Elizabeth Cauffman, Ph.D., Kimberly Gryglewicz, Ph.D. Afsaneh Razi, Heidi Hartikainen Ph.D., Zainab Agha, Neeraj Chatlani, Seunghyun Kim

Key Information: The following is a short summary of this study to help you decide whether or not to be a part of this study. More detailed information is listed later on in this form.

Why am I being invited to take part in a research study?

You are being invited take part in a research study because you are a young adult between the ages of 18-21 years old, have an active Instagram account, have received at least 2 direct message conversations from someone that made you or someone else feel uncomfortable or unsafe, and are willing to share your Instagram data for the purpose of research. Active is defined as having an Instagram account for the time period specified below and exchanging direct messages with at least 15 people.

In this research study, we want to know about your online experiences when you were a teen (ages 13-17), so we need to ensure that you had an active Instagram account during this timeframe. Therefore, depending on your age, you need to have had an Instagram account for time period specified below:

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

To be eligible, you must also currently reside in the United States and speak fluent English. An uncomfortable or unsafe interaction may include:

- **Nudity/porn**: Photos or videos of a nude or partially nude person.
- **Sexual messages or Solicitations**: Someone sent me a sexual message ("Sexting"). Someone asked me to send them a sexual message, revealing, or naked photo of myself.
- **Harassment**: Messages that contain credible threats, targets to degrade or shame someone, contains personal information to blackmail or harass someone, threats 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 or photos or videos of extreme violence, or encourage violence or attacks anyone based on their religious, ethnic or sexual background
- **Sale or promotion of illegal activities:** Messages contain promoting the use, sell, or distributing illegal material such as drugs.
- **Self-injury:** Messages encouraging or promoting self-injury, which includes suicide, cutting and eating disorders.
- **Other:** Other situations that could potentially lead to emotional or physical harm.

Why is this research being done?

As teens engage with others on social media, sometimes they encounter situations that make them feel uncomfortable or unsafe. We want to better understand these types of experiences teens encounter on social media (specifically on Instagram), so that we can design interventions that help teens feel safer and empowered online.

How long will the research last and what will I need to do?

If you sign this consent form, you will be enrolled in the study (as described in the "Detailed Information" section). You will be asked to download your Instagram data, take a web-based survey, upload your Instagram data, and take a survey based on your Instagram experiences. The entire process should take from 1 hour to 3 days to complete depending on how long it takes Instagram to prepare your data file for download. You will receive email reminders to complete each step of this process should you need to leave the study and come back later.

More detailed information about the study procedures can be found under ***"What happens if I say yes, I want to be in this research?"***

Is there any way being in this study could be bad for me?

The risks to participants are minimal and do not exceed the risks associated with activities found in daily life. However, the online survey includes questions about some sensitive topics related to social media use, such as cyberbullying, unwanted sexual experiences, mental health including depression and self-harm, and risky behaviors. Some people may become anxious or upset when answering questions about their behaviors, well-being or views. If you believe you need counseling, please contact 1-877-SAMHSA7 (1-877-726-4727), or refer to the resources provided in the "Help Resources" section available at all times during the study.

Will being in this study help me in any way?

Possible benefits include that participating in this study may increase your awareness of your social media activities.

What happens if I do not want to be in this research?

Participation in research is completely voluntary. You can decide to participate or not participate.

Detailed Information: The following is more detailed information about this study in addition to the information listed above.

What should I know about a research study?

- Someone could explain this research study to you.
- Whether or not you take part is up to you.
- You can choose not to take part.
- You can agree to take part and later change your mind.
- Your decision will not be held against you.
- You can ask all the questions you want before you decide.

Who can I talk to?

If you have questions, concerns, or complaints, or think the research has hurt you, talk to the research team by contacting Dr. Pamela Wisniewski at pamwis@ucf.edu.

This research has been reviewed and approved by an Institutional Review Board ("IRB"). You may talk to them at 407-823-2901or irb@ucf.edu if:

- Your questions, concerns, or complaints are not being answered by the research team.
- You cannot reach the research team.
- You want to talk to someone besides the research team.
- You have questions about your rights as a research subject.
- You want to get information or provide input about this research.

How many people will be studied?

For Part 1, we expect to enroll up to 1500 participants. Out of these, up to 75 teens or young adults will be recruited to participate in Part 2 of the study.

What happens if I say yes, I want to be in this research?

- After agreeing to participate in this study, you will be given instructions on how to download your Instagram data file. This process will take no more than 5 minutes to complete. However, it may take Instagram up to 48 hours to fulfill the request. You will later be asked to upload this file to our system. This file includes your Instagram posts, direct messages, and photos.
- Next, you will be asked to complete a web-based survey about your social media usage and personal experiences which includes questions about social media use, cyberbullying, unwanted sexual experiences, mental health including depression and self-harm, and risky behaviors. This survey will take approximately 30-60 minutes to complete.
- You will receive an email from Instagram to download your data file within 48 hours of your initial request. Once downloaded, you will be asked to upload this file to our secure system. This process should take approximately 5-30 minutes to complete depending on the file size. **Please do not upload any Instagram file that may contain any visual depiction of sexually explicit conduct involving a minor**

(persons less than 18 years old). Federal law requires us to report child pornography that we find in the data to the proper authorities.

- Once the Instagram data has been uploaded, you will be asked to identify direct messages that made you feel uncomfortable or unsafe in the past. We will ask you to answer questions to help us better understand these experiences. This survey should take approximately 30-90 minutes to complete.
- In case you close the browser you will receive an email with a link to continue and will receive reminder emails every 48 hours to continue. Participants can unsubscribe from emails at any time by hitting unsubscribe link in the emails. Unsubscribing from emails doesn't affect participation in the study.
- You may receive an offer to participate in phase 2 of the study which includes 30 minutes interview session that includes questions about how you felt participating in phase 1 of the study. If you receive this offer, you may choose to opt into the study, in which case the researchers will contact you in order to begin phase 2. If you choose not to participate, you may opt out, in which case you will not be contacted.

What happens if I say yes, but I change my mind later?

You can choose to leave the study at any time, and it will not be held against you. If you decide to leave the study before you finish, you will not receive any compensation. If you would like to completely withdraw from the study and not have the data collected used for the research, please email Dr. Pamela Wisniewski at pamwis@ucf.edu to make this request. If you partially complete the study but do not withdraw, we will use the data you have already shared.

What happens to the information collected for the research?

The information collected for this study will be kept confidential to the extent permitted by law. We take your privacy seriously and will treat your data with the utmost care. The Instagram data collected for this study may include personally identifiable information (e.g., names, phone numbers, addresses, etc.). However, no results will be published from this study that would allow you to be personally identified by the information disclosed. We will follow best practices to ensure that all data is encrypted and stored securely. Efforts will be made to limit the use and disclosure of your personal information included in this research study to people who have a need to review this information. However, we **cannot** promise complete secrecy. Organizations that may inspect and copy your information include the IRB and other representatives of this organization to ensure the ethical compliance of this research procedure is being upheld. As the research sponsor, the National Science Foundation may also inspect the data.

If data collected from this study strongly indicates that you are at serious risk of physical injury, sexual abuse, mental injury, or physical neglect we, as mandated reporters, are required to report these types of imminent risks to the proper authorities. However, the data will not be screened directly for such risks and will only be reported in the case that an imminent risk is found. The content of the Instagram data file may include information shared with you by other users. This data will be treated as part of your own data and will be under the same requirements for mandated reporting as your own

data.

All data collected must be retained for a minimum of five years per Florida statute. De-identified data must be retained for a minimum of ten years per the National Science Foundation's requirement for data retention for funded research. **These data retention laws apply even in the case that you or your parent choose to withdraw from the study.**

The data collected for this study will not be shared publicly but may be shared with a limited group of researchers who are working closely with the co-investigators. Those researchers will be required to comply with all assurances and regulations set forth by UCF's IRB. Portions of the social media data may also be shared with researchers requesting third-party access. The researchers will remove Instagram usernames and participants' contact information (including participants' names and email addresses) prior to sharing the data. To gain third-party access, researchers must show an established record of relevant, published research to validate why they should have access to this data, receive IRB approval from their home institutions, and sign a binding data confidentiality agreement with the primary investigator. Under no circumstances will third-parties be allowed to re-distribute any data collected from this study.

If you want your information to only be used for this research study, please request that by emailing Dr. Pamela Wisniewski at pamwis@ucf.edu within 2 weeks after you complete the study.

What else do I need to know?

If you complete this research study data upload and survey and your data is verified to match the study inclusion criteria, we will pay you a $50 Amazon gift card for your time and effort in completing this study. Gift cards will be sent via email to the address you specify after agreeing to this consent form. You need to complete the study and provide your Instagram data in other to be eligible to receive the gift card. If you exit the study after completing the survey but do not provide your Instagram data, you will not receive the gift card. Gift cards will be sent within 60 days of completing the study once the data has been verified based on the pre-screening eligibility requirements of this study. If your data is not found to match the study inclusion criteria, you will be contacted to be told why your data did not meet the inclusion criteria.

This research is covered by a Certificate of Confidentiality from the National Institutes of Health. This means that the researchers cannot release or use information, documents, or samples that may identify you in any action or suit unless you say it is okay. They also cannot provide them as evidence unless you have agreed. This protection includes federal, state, or local civil, criminal, administrative, legislative, or other proceedings. An example would be a court subpoena.

There are some important things that you need to know. The Certificate DOES NOT stop reporting that federal, state or local laws require. Some examples are laws that require reporting of child or elder abuse, child pornography, some communicable diseases, and threats to harm yourself or others. The Certificate CANNOT BE USED to stop a sponsoring United States federal or state government agency from checking records or evaluating programs. The Certificate also DOES NOT prevent your information from being used for other research if allowed by federal regulations.

Researchers may release information about you when you say it is okay. For example, you may give them permission to release information to insurers, medical providers or any other persons not connected with the research. The Certificate of Confidentiality does not stop you from willingly releasing information about your involvement in this research. It also does not prevent you from having access to your own information. The content shared with the participant by other users that is included in their Instagram data file, such as direct messages, will be treated as participant's data and will be under the same requirements for mandated reporting as the participant's own data.

This research is being funded by the National Science Foundation. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the U.S. National Science Foundation.

After signing this form, choose "Print" from your browser menu or use the "Ctrl+P" keyboard shortcut to print this document. You may also save as a PDF for your records.

Thank you for being a part of our research study!

APPENDIX D: SURVEY MEASURES(STUDY 2)

Social Media and Youth Survey

Thank you for agreeing to participate in our research study. This survey should take you less than 25 minutes to complete. We will ask you about your social media use, online experiences, and personal experiences. Please answer the questions honestly. There are no wrong answers.

Social Media Use Questions

1. **How often do you use your Instagram account?**
    a. Less than once a month
    b. Less than once a week
    c. Once or twice a week
    d. Every day or almost every day
    e. Several times a day
    f. Several times an hour
    g. Never

If Never, they are ineligible to participate in the study.

2. **Which social media platform do you use MOST often?**
    a. Instagram
    b. Snapchat
    c. Facebook
    d. Twitter
    e. Tumblr
    f. Reddit
    g. YouTube
    h. Other, please specify

3. **Do you have more than one Instagram account?**
    a. Yes
    b. No

If Yes, ask additional question.

3a. Please explain why you have multiple Instagram accounts. Also, tell us which one you plan to use for this research and why. We would like you to use your primary personal account that you use most frequently.

[Open text box, required if asked]

Please answer the following questions based on your primary Instagram account that you are using for this study.

4. **Do you have Instagram friends who you have never met in person?**
    a. Yes
    b. No

5. **With whom have you exchanged direct messages with on Instagram in the past? (Check all that Apply)**
    a. Family
    b. Romantic Partner
    c. Friends
    d. Acquaintances
    e. Strangers
    f. Other, please specify

6. **How often do you use disappearing messages on Instagram?**
    a. Never
    b. Rarely
    c. Quite Often
    d. All the Time
    e. I don't know what disappearing messages are

Please indicate how strongly you disagree or agree to the following statements.

Facebook Intensity scale (revised)

[1-Strongly Disagree, Disagree, Neutral, Agree, 5-Strongly Agree]

7. Instagram is part of my everyday activity
8. I am proud to tell people I'm on Instagram
9. Instagram has become part of my daily routine
10. I feel out of touch when I haven't logged onto Instagram for a while
11. I feel I am part of the Instagram community
12. I would be sorry if Instagram shut down

References

Ellison, N.B. et al. 2007. The Benefits of Facebook "Friends:" Social Capital and College Students' Use of Online Social Network Sites. Journal of Computer-Mediated Communication. 12, 4 (Jul. 2007), 1143–1168. DOI:https://doi.org/10.1111/j.1083-6101.2007.00367.x.

Social Capital – Bonding scale (revised)

[1-Strongly Disagree, Disagree, Neutral, Agree, 5-Strongly Agree]

13. There is someone on Instagram I can turn to for advice about making very important decisions.
14. There are several people on Instagram I trust to help solve my problems.
15. The people I interact with on Instagram would put their reputation on the line for me.
16. I do not know people on Instagram well enough to get them to do anything important. (REVERSE CODED)
17. When I feel lonely, there are several people on Instagram I can talk to.

Social Capital – Bridging scale (revised)

[1-Strongly Disagree, Disagree, Neutral, Agree, 5-Strongly Agree]

18. Interacting with people on Instagram makes me feel connected to the bigger picture.
19. Talking with people on Instagram makes me curious about other places in the world.
20. Interacting with people on Instagram makes me want to try new things.
21. Interacting with people on Instagram makes me interested in things that happen outside of my town.
22. Interacting with people on Instagram makes me feel like part of a larger community.

References

Ellison, N.B. et al. 2007. The Benefits of Facebook "Friends:" Social Capital and College Students' Use of Online Social Network Sites. Journal of Computer-Mediated Communication. 12, 4 (Jul. 2007), 1143–1168. DOI:https://doi.org/10.1111/j.1083-6101.2007.00367.x.

Please indicate how frequently each of the statements below applied to you IN THE PAST YEAR.

[1=Not at all, 2=One to a few times this past year, 3= A few times a month, 4= A few times a week, 5= Almost every day]

Social Media Disorder Scale

23. Regularly found that you can't think of anything else but the moment that you will be able to use social media again
24. Regularly felt dissatisfied because you wanted to spend more time on social media
25. Often felt bad when you could not use social media
26. Tried to spend less time on social media, but failed
27. Regularly neglected other activities (e.g. hobbies, sport) because you wanted to use social media
28. Regularly had arguments with others because of your social media use
29. Regularly lied to your parents or friends about the amount of time you spend on social media?
30. Often used social media to escape from negative feelings
31. Had serious conflict with your parents, brother(s) or sister(s) because of your social media use

References

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: 478–487. https://doi.org/10.1016/j.chb.2016.03.038

Online Experience Questions

Next, we will ask you some questions regarding potentially negative experiences you have had on Instagram. Again, there are no wrong answers. We want to understand your personal experiences.

Cyber-victimization (CAV-V) (Shapka)

Please indicate how frequently each of the statements below applied to you IN THE PAST YEAR.

[1=Not at all, 2=One to a few times this past year, 3= A few times a month, 4= A few times a week, 5= Almost every day]

32. Had something embarrassing or mean posted or re-posted about you on Instagram.
33. Received a hurtful message from someone on Instagram.
34. Had an embarrassing photo or video of you posted or re-posted on Instagram that you didn't want others to see.
35. Had hurtful comments made on Instagram about an online photo or video of you.
36. Been purposely excluded by others on Instagram.
37. Had something personal posted or re-posted about you on Instagram that you didn't want others to know.
38. Had gossip or rumors spread about you on Instagram.
39. Received hurtful comments or messages about your race or ethnicity on Instagram.
40. Received hurtful comments or messages about your perceived sexual orientation on Instagram.
41. Received hurtful comments about your perceived sexual behaviors on Instagram.
42. Received a sexual message from somebody on Instagram who was trying to be mean to you or to embarrass you.
43. Had sexual content (photos or jokes) sent to you from somebody on Instagram who was trying to be mean to you or embarrass you.

Cyber-aggression Perpetration (CAV-P) (Shapka)

Please indicate how frequently each of the statements below applied to you IN THE PAST YEAR.

[1=Not at all, 2=One to a few times this past year, 3= A few times a month, 4= A few times a week, 5= Almost every day]

44. Posted or re-posted something embarrassing or mean about another person on Instagram.
45. Sent or forwarded a hurtful message to someone on Instagram.
46. Posted or re-posted an embarrassing photo or video of someone on Instagram that he or she did not want others to see.
47. Posted or texted a hurtful comment on Instagram about a photo or video of somebody else.
48. Posted or sent messages on Instagram to purposely exclude a certain person or group of people.
49. Posted or re-posted something private on Instagram about another person that he or she did not want others to know.
50. Used Instagram to spread rumors or gossip about someone.
51. Made hurtful comments about somebody's race or ethnicity on Instagram.
52. Made hurtful comments about somebody's perceived sexual orientation on Instagram.
53. Made hurtful comments about somebody's perceived sexual behaviors on Instagram.

54. Said something sexual to somebody else on Instagram to embarrass them or to be mean.
55. Sent sexual content (photos or jokes) to somebody else on Instagram to embarrass them or to be mean.

References

[12] 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: 10–17. https://doi.org/10.1016/j.chb.2016.12.015

Cyberbullying Victimization and perpetuation scale (Doane- 54 citations)

Please indicate how frequently each of the statements below applied to you IN THE PAST YEAR.

[1=Not at all, 2=One to a few times this past year, 3= A few times a month, 4= A few times a week, 5= Almost every day]

Deception

56. Someone has pretended to be someone else while talking to you on Instagram.
57. Someone has lied about themselves to you on Instagram.
58. You shared personal information with someone on Instagram and then later found the person was not who you thought they were.

Reference

Doane, A. N., Kelley, M. L., Chiang, E. S., & Padilla, M. A. (2013). Development of the Cyberbullying Experiences Survey. Emerging Adulthood, 1(3), 207–218. https://doi.org/10.1177/2167696813479584

YISS Unwanted Online Experiences (sexual solicitation, unwanted exposure to sexual material)

Please indicate how frequently each of the statements below applied to you IN THE PAST YEAR. Unwanted sexual solicitations are defined as requests to engage in sexual activities or sexual talk or to give personal sexual information that were unwanted or made by a person 5 or more years older, whether wanted or not.

[1=Not at all, 2=One to a few times a year, 3= A few times a month, 4= A few times a week, 5= Almost every day]

59. Someone tried to get me to talk on Instagram about sex when I did not want to.
60. Someone on Instagram asked me for sexual information about myself when I did not want to answer such questions. (Very personal questions, like what your body looks like or sexual things you have done)
61. Someone on Instagram asked me to do something sexual that I did not want to do.

Please indicate how frequently each of the statements below applied to you IN THE PAST YEAR. Unwanted exposure to pornography is defined as being exposed to pictures or videos of naked people or people having sex without seeking or expecting it.

[1=Not at all, 2=One to a few times this past year, 3= A few times a month, 4= A few times a week, 5= Almost every day]

62. My Instagram feed showed me pictures of naked people or people having sex when I did not want to see such content.
63. Someone on Instagram sent me a direct message of naked people or people having sex that I did not want to see.

Youth Produced Sexual Images (Sexting) (YISS)

Please indicate how frequently each of the statements below applied to you IN THE PAST YEAR. By 'nearly nude' we mean pictures in one's underwear or partially undressed.

64. I have shared a nude or nearly nude picture or video of myself on Instagram.
65. Someone else shared a nude or nearly nude picture or video of me on Instagram.
66. Were there any negative experiences you had on Instagram that made you feel uncomfortable or unsafe that were not covered in the questions above? If so, please describe in detail, so that we can better understand your experiences.

[Open textbox, not required]

Reference

Mitchell, K.J. & Jones, L.M. (2012). Youth Internet Safety Study: Methodology Report. Durham, NH: Crimes against Children Research Center.

Personal Experiences and Demographic Questions

Next, we will ask questions related to your personal experiences and demographic information. Again, there are no wrong answers. We want to better idea about the people who are participating in our study.

The Short Warwick-Edinburgh Mental Well-being Scale (SWEMWBS)

Below are some statements about your feelings and thoughts. Please select the answer that best describes your experience of each statement over the last week.

[1-None of the Time, Rarely, Some of the Time, Often, 5-All of the Time]

67. I've been feeling optimistic about the future.
68. I've been feeling useful.
69. I've been feeling relaxed.
70. I've been dealing with problems well.
71. I've been thinking clearly.
72. I've been feeling close to other people.

73. I've been able to make up my own mind about things.

Reference

Tennant, R., Hiller, L., Fishwick, R. et al. The Warwick-Edinburgh Mental Well-being Scale (WEMWBS): development and UK validation. Health Qual Life Outcomes 5, 63 (2007). https://doi.org/10.1186/1477-7525-5-63

ULS-8 Loneliness Scale

[1-None of the Time, Rarely, Some of the Time, Often, 5-All of the Time]

74. I lack companionship.
75. There is no one I can turn to.
76. I am an outgoing person.
77. I feel left out.
78. I feel isolated from others.
79. I can find companionship when I want it.
80. I am unhappy being so withdrawn.
81. People are around me but not with me.

Reference

Ron D. Hays & M. Robin DiMatteo (1987) A Short-Form Measure of Loneliness, Journal of Personality Assessment, 51:1, 69-81, DOI: 10.1207/s15327752jpa5101_6

Depression (PHQ-9 Scale)

Over the past 2 weeks, how often have you been bothered by any of the following problems?

[1-None of the Time, Rarely, Some of the Time, Often, 5-All of the Time]

82. Little interest or pleasure in doing things
83. Feeling down, depressed or hopeless
84. Trouble falling asleep, staying asleep, or sleeping too much
85. Feeling tired or having little energy
86. Poor appetite or overeating
87. Feeling bad about yourself – or that you're a failure or have let yourself or your family down
88. Trouble concentrating on things, such as reading the newspaper or watching television
89. Moving or speaking so slowly that other people could have noticed. Or, the opposite being so fidgety or restless that you have been moving around a lot more than usual
90. Thoughts that you would be better off dead or of hurting yourself in some way

Reference

Kurt Kroenke, MD; Robert L Spitzer, MD, The PHQ-9: A New Depression Diagnostic and Severity Measure, Psychiatric Annals. 2002;32(9):509-515https://doi.org/10.3928/0048-5713-20020901-06

INVENTORY OF STATEMENTS ABOUT SELF-INJURY (ISAS) FIRST QUESTION

This question asks about a variety of self-harm behaviors. Please only endorse a behavior if you have done it intentionally (i.e., on purpose) and without intent (i.e., not for suicidal reasons).

Please estimate how often in your life you have intentionally (i.e., on purpose) performed each type of non-suicidal self-harm:

[1- None of the Time, Rarely, Some of the Time, Often, 5-All of the Time]

91. Cutting
92. Severe Scratching
93. Biting
94. Banging or Hitting Self
95. Burning
96. Interfering w/ Wound Healing (e.g., picking scabs)
97. Carving
98. Rubbing Skin Against Rough Surface
99. Pinching
100. Sticking Self w/ Needles
101. Pulling Hair
102. Swallowing Dangerous Substances
103. Other

Reference

Klonsky, E.D. & Glenn, C.R. J, Assessing the Functions of Non-suicidal Self-injury: Psychometric Properties of the Inventory of Statements About Self-injury (ISAS), J Psychopathol Behav Assess. 2009 Sep;31(3):215-219. doi: 10.1007/s10862-008-9107-z

Risky Behavior Questionnaire for Adolescents (RBQ-A)

In this questionnaire we are interested in whether certain events have happened to you. Please indicate how often the following events have happened to you.

Please use the following scale: 0 ¼ Never; 1 ¼ Almost never (1 time per month); 2 ¼ Sometimes (2e4 times per month); 3 ¼ Almost always (2e3 times per week); 4 ¼ Always (4 or more times per week).

104. Have you destroyed property (other than your own)?
105. Have you been unfaithful to your boyfriend or girlfriend?
106. Have you been in a physical fight?
107. Have you bullied, threatened, or intimidated a peer(s)?
108. Have you been binge drinking and/or drinking to get drunk?
109. Have you used illegal drugs?
110. Have you sold illegal drugs?
111. Have you skipped class (or entire days of school)?

112. Have you cheated or plagiarized?
113. Have you shoplifted?
114. Have you stolen money?
115. Have you had unsafe sex?
116. Have you verbally harassed someone?
117. Have you made attempts to cut or burn yourself?
118. Have you purged or binged?
119. Have you gambled?
120. Have you lied to your family members (e.g., grandparents, parents, siblings)?
121. Have you driven (a bicycle, a moped, and/or a car) recklessly (e.g., at fast speeds, under the influence of a substance)?
122. Have you used cigarettes?
123. Have you engaged in acts of revenge?

Reference

Randy P. Auerbach, Casey K. Gardiner, Moving beyond the trait conceptualization of self-esteem: The prospective effect of impulsiveness, coping, and risky behavior engagement, Behaviour Research and Therapy, Volume 50, Issue 10, 2012, Pages 596-603, ISSN 0005-7967, https://doi.org/10.1016/j.brat.2012.06.002.

Demographic Information

Next you will be asked some questions about your background and demographics:

124. Please select your gender.

a. Male
b. Female
c. Non-Binary
d. Prefer to self-identify _____

125. Please select your current age.

a. 13
b. 14
c. 15
d. 16
e. 17
f. 18
g. 19
h. 20
i. 21

126. Please select the city and state in which you live.

[Drop down of U.S. states and territories]

[State drop down would dynamically populate based on the state selected]

127. Please select your race. Select all that apply.

a. White/Caucasian
b. Black/African-American
c. Hispanic/Latino
d. Asian or Pacific Islander
e. American Indian/Alaska Native
f. Prefer to Self-Identify _____

128. What is your sexual orientation?

a. Heterosexual or straight
b. Homosexual or gay
c. Bisexual
d. Prefer to self-identify _____

129. What is your current relationship status?

a. Single
b. Dating (nonexclusive)
c. Serious relationship (exclusive relationship)
d. Cohabitating (living together)

e. Married
f. Prefer to self-identify _____

130. What best describes who were your primary caregiver(s) when you were a teenager?

a. Mother and Father
b. Mother and Stepfather
c. Father and Stepmother
d. Mother only
e. Father only
f. Adopted or Foster Parents
g. Adopted or Foster Mother only
h. Adopted or Foster Father only
i. Grandmother and/or Grandfather
j. Other, please specify _____

Thank you. You have completed this part of the study. Next, you will be asked to upload your Instagram data to continue the second part of this study.

Please check the email address you provided for your Instagram data download. Have you received an email with a link to download your data?

- Yes
- No

If the answer is yes:

Direct them to instructions to download and upload the data

If the answer if no:

If you haven't received an email from Instagram yet with a link to download your data, you can close this window at this time. We will email you a link to upload your Instagram data and proceed with the study after you have been able to download your data from Instagram.

LIST OF REFERENCES

[1] Fact Sheet: the Obama Administration Announces Efforts to Combat Human Trafficking at Home and Abroad, Sept. 2012. URL https://obamawhitehouse.archives.gov/the-press-office/2012/09/25/fact-sheet-obama-administration-announces-efforts-combat-\human-trafficki.

[2] 105 leading social networks worldwide. practical ecommerce. https://www.practicalecommerce.com/105-leading-social-networks-worldwide, 2017.

[3] Nsf award search: Award#1827700 - pfi-rp: A multi-disciplinary approach to detecting adolescent online risks. https://nsf.gov/awardsearch/showAward?AWD_ID=1827700&HistoricalAwards=false, 2018.

[4] Nsf award search: Award#1928627 - fw-htf-rm: Collaborative research: Augmenting social media content moderation. https://www.nsf.gov/awardsearch/showAward?AWD_ID=1928627&HistoricalAwards=false, 2019.

[5] Nsf award search: Award#1764089 - chs: Medium: Scaling qualitative inductive analysis through computational methods. https://www.nsf.gov/awardsearch/showAward?AWD_ID=1764089&HistoricalAwards=false, 2019.

[6] National center for missing and exploited children, 2020. URL https://www.missingkids.org/footer/media/keyfacts.

[7] T. V. P. Act. Victims of trafficking and violence protection act of 2000. *United States*, 2000.

[8] A. Adadi and M. Berrada. Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access*, 6:52138–52160, 2018.

[9] Z. Agha, N. Chatlani, A. Razi, and Pamela Wisniewski. Towards Conducting Responsible Research with Teens and Parents regarding Online Risks. In *CHI 2020 Extended Abstracts*. doi: https://doi.org/10.1145/3334480.3383073.

[10] Z. Agha, R. Ghaiumy Anaraky, K. Badillo-Urquiola, B. McHugh, and P. Wisniewski. 'just-in-time'parenting: A two-month examination of the bi-directional influences between parental mediation and adolescent online risk exposure. In *International Conference on Human-Computer Interaction*, pages 261–280. Springer, 2021.

[11] H. Al Kuwatly, M. Wich, and G. Groh. Identifying and measuring annotator bias based on annotators' demographic characteristics. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 184–190, 2020.

[12] S. Ali, A. Razi, S. Kim, A. Alsoubai, J. Gracie, M. De Choudhury, P. J. Wisniewski, and G. Stringhini. Understanding the Digital Lives of Youth: Analyzing Media Shared within Safe Versus Unsafe Private Conversations on Instagram, 2022. URL `https://camps.aptaracorp.com/ACM{_}PMS/PMS/ACM/CHI22/151/591a33b1-4cc3-11ec-b613-166a08e17233/OUT/chi22-151.html`.

[13] M. H. Aljanahi. "you could say i'm a hardcore fan of dragon ball z": Affinity spaces, multiliteracies, and the negotiation of identity. *Literacy Research and Instruction*, 58(1): 31–48, Sept. 2018. doi: 10.1080/19388071.2018.1520940. URL `https://doi.org/10.1080/19388071.2018.1520940`.

[14] A. Alkhatib and M. Bernstein. Street-level algorithms: A theory at the gaps between policy and decisions. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2019.

[15] F. Amuchi, A. Al-Nemrat, M. Alazab, and R. Layton. Identifying cyber predators through forensic authorship analysis of chat logs. In *2012 Third Cybercrime and Trustworthy Computing Workshop*, pages 28–37. IEEE, 2012.

[16] N. Andalibi and A. Forte. Social computing researchers as vulnerable populations. 01 2015.

[17] N. Andalibi and A. Forte. Social computing researchers as vulnerable populations. In *ACM Conference on Computer Supported Cooperative Work & Social Computing Workshop on Ethics for Studying Sociotechnical Systems in a Big Data World*, 2015.

[18] N. Andalibi and A. Forte. Announcing pregnancy loss on facebook: A decision-making framework for stigmatized disclosures on identified social network sites. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–14, 2018.

[19] N. Andalibi, O. L. Haimson, M. D. Choudhury, and A. Forte. Understanding social media disclosures of sexual abuse through the lenses of support seeking and anonymity. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, May 2016. doi: 10.1145/2858036.2858096. URL `https://doi.org/10.1145/2858036.2858096`.

[20] N. Andalibi, O. L. Haimson, M. De Choudhury, and A. Forte. Understanding Social Media Disclosures of Sexual Abuse Through the Lenses of Support Seeking and Anonymity. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16*, pages 3906–3918, Santa Clara, California, USA, 2016. ACM Press. ISBN 978-1-4503-3362-7. doi: 10.1145/2858036.2858096. URL `http://dl.acm.org/citation.cfm?doid=2858036.2858096`.

[21] N. Andalibi, O. L. Haimson, M. D. Choudhury, and A. Forte. Social support, reciprocity, and anonymity in responses to sexual abuse disclosures on social media. *ACM Transactions*

on Computer-Human Interaction, 25(5):1–35, Oct. 2018. doi: 10.1145/3234942. URL https://doi.org/10.1145/3234942.

[22] M. Anderson and J. Jiang. Teens, social media and technology 2018. http://www.pewinternet.org/2018/05/31/teens-social-media-technology-2018/, 2018. [Pew Research Center. Retrieved September 22, 2018 from].

[23] M. Anderson and J. Jiang. Teens, Social Media & Technology 2018 | Pew Research Center, May 2018. URL http://www.pewinternet.org/2018/05/31/teens-social-media-technology-2018/.

[24] P. Anderson, Z. Zuo, L. Yang, and Y. Qu. An Intelligent Online Grooming Detection System Using AI Technologies. In *2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–6, June 2019. doi: 10.1109/FUZZ-IEEE.2019.8858973. ISSN: 1544-5615.

[25] A. N. Antle. The ethics of doing research with vulnerable populations. *interactions*, 24 (6):74–77, Oct. 2017. ISSN 10725520. doi: 10.1145/3137107. URL http://dl.acm.org/citation.cfm?doid=3155029.3137107.

[26] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115, 2020.

[27] M. Ashcroft, L. Kaati, and M. Meyer. A Step Towards Detecting Online Grooming – Identifying Adults Pretending to be Children. In *2015 European Intelligence and Security Informatics Conference*, pages 98–104, Sept. 2015. doi: 10.1109/EISIC.2015.41.

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

[29] K. Badillo-Urquiola, Z. Shea, Z. Agha, I. Lediaeva, and P. Wisniewski. 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):1–46, 2021.

[30] K. A. Badillo-Urquiola, X. Page, and P. Wisniewski. Risk vs. restriction: The tension between providing a sense of normalcy and keeping foster teens safe online. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019.

[31] T. Baltrušaitis, C. Ahuja, and L.-P. Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.

[32] R. Barakat, S. Abufardeh, and K. Magel. Automated framework to improve users' awareness on online social networks. In *2016 IEEE International Conference on Electro Information Technology (EIT)*, pages 0428–0433. IEEE, 2016.

[33] C. Barber and S. Bettez. Deconstructing the online grooming of youth: Toward improved information systems for detection of online sexual predators. Jan. 2014.

[34] R. Barreira, V. Pinheiro, and V. Furtado. A framework for digital forensics analysis based on semantic role labeling. In *2017 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 66–71. IEEE, 2017.

[35] E. P. Baumer. Toward human-centered algorithm design. *Big Data & Society*, 4(2): 2053951717718854, Dec. 2017. ISSN 2053-9517. doi: 10.1177/2053951717718854. URL `https://doi.org/10.1177/2053951717718854`.

[36] S. E. Baumgartner, P. M. Valkenburg, and J. Peter. Unwanted online sexual solicitation and risky sexual online behavior across the lifespan. *Journal of Applied Developmental Psychology*, 31(6):439–447, Nov. 2010. ISSN 0193-3973. doi: 10.1016/j.appdev.2010.07.005. URL http://www.sciencedirect.com/science/article/pii/S0193397310000857.

[37] M. Bayer, M.-A. Kaufhold, and C. Reuter. A survey on data augmentation for text classification. *arXiv preprint arXiv:2107.03158*, 2021.

[38] V. Bellotti and K. Edwards. Intelligibility and accountability: human considerations in context-aware systems. *Human–Computer Interaction*, 16(2-4):193–212, 2001.

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

[40] R. Benjamin. Assessing risk, automating racism. *Science*, 366(6464):421–422, 2019.

[41] W. E. Bijker, T. P. Hughes, T. Pinch, et al. The social construction of technological systems, 1987.

[42] P. J. Black, M. Wollis, M. Woodworth, and J. T. Hancock. A linguistic analysis of grooming strategies of online child sex offenders: Implications for our understanding of predatory sexual behavior in an increasingly computer-mediated world. *Child Abuse & Neglect*, 44: 140–149, June 2015. ISSN 0145-2134. doi: 10.1016/j.chiabu.2014.12.004. URL http://www.sciencedirect.com/science/article/pii/S0145213414004360.

[43] L. Blackwell, E. Gardiner, and S. Schoenebeck. Managing expectations: Technology tensions among parents and teens. *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, 2016.

[44] L. Blackwell, M. Handel, S. T. Roberts, A. Bruckman, and K. Voll. Understanding "bad actors" online. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, Apr. 2018. doi: 10.1145/3170427.3170610. URL https://doi.org/10.1145/3170427.3170610.

[45] D. Bogdanova, P. Rosso, and T. Solorio. On the Impact of Sentiment and Emotion Based Features in Detecting Online Sexual Predators. In *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*, WASSA '12, pages 110–118, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=2392963.2392986.

[46] D. Bogdanova, P. Rosso, and T. Solorio. Modelling Fixated Discourse in Chats with Cyberpedophiles. In *Proceedings of the Workshop on Computational Approaches to Deception Detection*, EACL 2012, pages 86–90, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=2388616.2388629. event-place: Avignon, France.

[47] D. Bogdanova, P. Rosso, and T. Solorio. Exploring high-level features for detecting cyberpedophilia. *Computer speech & language*, 28(1):108–120, 2014.

[48] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in Neural Information Processing Systems*, 2016.

[49] P. R. Borj and P. Bours. Predatory Conversation Detection. In *2019 International Conference on Cyber Security for Emerging Technologies (CSET)*, pages 1–6, Oct. 2019. doi: 10.1109/CSET.2019.8904885. ISSN: null.

[50] P. Bourdieu and L. J. Wacquant. *An invitation to reflexive sociology*. University of Chicago press, 1992.

[51] P. Bours and H. Kulsrud. Detection of Cyber Grooming in Online Conversation. In *2019 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6, Dec. 2019. doi: 10.1109/WIFS47025.2019.9035090. ISSN: 2157-4774.

[52] D. Boyd. It's complicated: The social lives of networked teens. 2015.

[53] V. Braun and V. Clarke. Thematic analysis. In *APA handbook of research methods in psychology, Vol 2: Research designs: Quantitative, qualitative, neuropsychological, and biological.*, pages 57–71. American Psychological Association, 2012. doi: 10.1037/13620-004. URL https://doi.org/10.1037/13620-004.

[54] V. Braun and V. Clarke. Thematic analysis. In *APA handbook of research methods in psychology, Vol 2: Research designs: Quantitative, qualitative, neuropsychological, and biological*, pages 57–71. American Psychological Association, Washington, DC, US, 2012. ISBN 978-1-4338-1005-3. doi: 10.1037/13620-004.

[55] M. Brenner, J. S. Gersen, M. Haley, M. Lin, A. Merchant, R. J. Millett, S. K. Sarkar, and D. Wegner. Constitutional dimensions of predictive algorithms in criminal justice. *Harv. CR-CLL Rev.*, 55:267, 2020.

[56] L. J. Broome, C. Izura, and J. Davies. A psycho-linguistic profile of online grooming conversations: A comparative study of prison and police staff considerations. *Child Abuse and Neglect*, 109:104647, 2020. ISSN 0145-2134. doi: https://doi.org/10.1016/j.chiabu.2020.104647. URL https://www.sciencedirect.com/science/article/pii/S0145213420303021.

[57] A. Bruckman. Studying the amateur artist: A perspective on disguising data collected in human subjects research on the internet. *Ethics and Information Technology*, 4(3):217–231, 2002.

[58] J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

[59] X. Caddle, A. Alsoubai, A. Razi, S. Kim, S. Ali, G. Stringhini, M. D. Choudhury, and P. Wisniewski. Instagram data donation: A case for partnering with social media platforms to protect adolescents online. In *ACM Conference on Human Factors in Computing Systems (CHI 2021)/Social Media as a Design and Research Site in HCI: Mapping Out Opportunities and Envisioning Future Uses Workshop*, 2021.

[60] X. V. Caddle, A. Razi, S. Kim, S. Ali, T. Popo, G. Stringhini, M. De Choudhury, and P. J. Wisniewski. *MOSafely: Building an Open-Source HCAI Community to Make the Internet a Safer Place for Youth*, page 315–318. Association for Computing Machinery, New York, NY, USA, 2021. ISBN 9781450384797. URL https://doi.org/10.1145/3462204.3481731.

[61] A. E. Cano, M. Fernandez, and H. Alani. Detecting Child Grooming Behaviour Patterns on Social Media. In L. M. Aiello and D. McFarland, editors, *Social Informatics: 6th International Conference, SocInfo 2014, Barcelona, Spain, November 11-13, 2014. Proceedings*, Lecture Notes in Computer Science, pages 412–427. Springer International Publishing, Cham, 2014. ISBN 978-3-319-13734-6. doi: 10.1007/978-3-319-13734-6_30. URL https://doi.org/10.1007/978-3-319-13734-6_30.

[62] S. Chancellor, N. Andalibi, L. Blackwell, D. Nemer, and W. Moncur. Sensitive research, practice and design in hci. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–8, 2019.

[63] S. Chancellor, E. P. S. Baumer, and M. De Choudhury. Who is the "human" in human-centered machine learning: The case of predicting mental health from social media. *Proc.*

ACM Hum.-Comput. Interact., 3(CSCW), Nov. 2019. doi: 10.1145/3359249. URL https://doi.org/10.1145/3359249.

[64] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[65] J. H. Chen and A. Verghese. Planning for the known unknown: Machine learning for human healthcare systems. *The American Journal of Bioethics*, 20(11):1–3, 2020.

[66] X.-W. Chen and X. Lin. Big data deep learning: challenges and perspectives. *IEEE access*, 2:514–525, 2014.

[67] Y. Cheong, A. K. Jensen, E. R. Guðnadóttir, B. Bae, and J. Togelius. Detecting Predatory Behavior in Game Chats. *IEEE Transactions on Computational Intelligence and AI in Games*, 7(3):220–232, Sept. 2015. ISSN 1943-068X. doi: 10.1109/TCIAIG.2015.2424932.

[68] M. M. Chiu, K. C. Seigfried-Spellar, and T. R. Ringenberg. Exploring detection of contact vs. fantasy online sexual offenders in chats with minors: Statistical discourse analysis of self-disclosure and emotion words. *Child abuse & neglect*, 81:128–138, 2018.

[69] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. Natural language processing (almost) from scratch. *Journal of machine learning research*, 12 (ARTICLE):2493–2537, 2011.

[70] J. M. Corbin and A. Strauss. Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative Sociology*, 13(1):3–21, 1990. doi: 10.1007/bf00988593. URL https://doi.org/10.1007/bf00988593.

[71] J. A. Dake, J. H. Price, L. Maziarz, and B. Ward. Prevalence and correlates of sexting behavior in adolescents. *American Journal of Sexuality Education*, 7(1):1–15, Jan. 2012. doi:

10.1080/15546128.2012.650959. URL https://doi.org/10.1080/15546128.2012.650959.

[72] M. De Choudhury and S. De. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *Eighth international AAAI conference on weblogs and social media*, 2014.

[73] P. De Santisteban and M. Gámez-Guadix. Prevalence and risk factors among minors for online sexual solicitations and interactions with adults. *J. Sex Research*, 55(7):939–950, 2018.

[74] Z. Dhouioui and J. Akaichi. Privacy Protection Protocol in Social Networks Based on Sexual Predators Detection. In *Proceedings of the International Conference on Internet of Things and Cloud Computing*, ICC '16, pages 63:1–63:6, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4063-2. doi: 10.1145/2896387.2896448. URL http://doi.acm.org/10.1145/2896387.2896448.

[75] V. Dickson-Swift, E. L. James, S. Kippen, and P. Liamputtong. Blurring boundaries in qualitative health research on sensitive topics. *Qualitative health research*, 16(6):853–871, 2006.

[76] V. Dickson-Swift, E. L. James, S. Kippen, and P. Liamputtong. Doing sensitive research: what challenges do qualitative researchers face? *Qualitative research*, 7(3):327–353, 2007.

[77] A. L. Dir, A. Coskunpinar, J. L. Steiner, and M. A. Cyders. Understanding differences in sexting behaviors across gender, relationship status, and sexual identity, and the role of expectancies in sexting. *Cyberpsychology, Behavior, and Social Networking*, 16(8):568–574, 2013.

[78] A. L. Dir, A. Coskunpinar, J. L. Steiner, and M. A. Cyders. Understanding Differences in

Sexting Behaviors Across Gender, Relationship Status, and Sexual Identity, and the Role of Expectancies in Sexting. *Cyberpsychology, Behavior, and Social Networking*, 16(8): 568–574, May 2013. ISSN 2152-2715. doi: 10.1089/cyber.2012.0545. URL https://www.liebertpub.com/doi/10.1089/cyber.2012.0545.

[79] A. N. Doane, M. L. Kelley, E. S. Chiang, and M. A. Padilla. Development of the cyberbullying experiences survey. *Emerging Adulthood*, 1(3):207–218, 2013.

[80] S. M. Doornwaard, M. A. Moreno, R. J. van den Eijnden, I. Vanwesenbeeck, and T. F. ter Bogt. Young adolescents' sexual and romantic reference displays on facebook. *Journal of Adolescent Health*, 55(4):535–541, Oct. 2014. doi: 10.1016/j.jadohealth.2014.04.002. URL https://doi.org/10.1016/j.jadohealth.2014.04.002.

[81] N. Döring. Consensual sexting among adolescents: Risk prevention through abstinence education or safer sexting? *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, 8(1), Mar. 2014. doi: 10.5817/cp2014-1-9. URL https://doi.org/10.5817/cp2014-1-9.

[82] M. Drouin and E. Tobin. Unwanted but consensual sexting among young adults: Relations with attachment and sexual motivations. *Computers in Human Behavior*, 31:412–418, 2014.

[83] K. F. Durkin and C. D. Bryant. "log on to sex": Some notes on the carnal computer and erotic cyberspace as an emerging research frontier. *Deviant Behavior*, 16(3):179–200, July 1995. doi: 10.1080/01639625.1995.9967998. URL https://doi.org/10.1080/01639625.1995.9967998.

[84] M. Ebrahimi and M. Ebrahimi. Automatic Identification of Online Predators in Chat Logs by Anomaly Detection and Deep Learning. Master's thesis, Concordia University, Apr. 2016. URL https://spectrum.library.concordia.ca/981404/.

[85] M. Ebrahimi, C. Y. Suen, and O. Ormandjieva. Detecting predatory conversations in social media by deep Convolutional Neural Networks. *Digital Investigation*, 18:33–49, Sept. 2016. ISSN 1742-2876. doi: 10.1016/j.diin.2016.07.001. URL http://www.sciencedirect.com/science/article/pii/S1742287616300731.

[86] M. Ebrahimi, C. Y. Suen, O. Ormandjieva, and A. Krzyzak. Recognizing Predatory Chat Documents using Semi-supervised Anomaly Detection, Feb. 2016. URL https://www.ingentaconnect.com/content/ist/ei/2016/00002016/00000017/art00012#.

[87] N. B. Ellison, C. Steinfield, and C. Lampe. The Benefits of Facebook "Friends:" Social Capital and College Students' Use of Online Social Network Sites. *Journal of Computer-Mediated Communication*, 12(4):1143–1168, July 2007. doi: 10.1111/j.1083-6101.2007.00367.x. URL https://academic.oup.com/jcmc/article/12/4/1143/4582961.

[88] S. Elo and H. Kyngäs. The qualitative content analysis process. *Journal of advanced nursing*, 62(1):107–115, 2008.

[89] P. Elzinga, K. E. Wolff, and J. Poelmans. Analyzing Chat Conversations of Pedophiles with Temporal Relational Semantic Systems. In *2012 European Intelligence and Security Informatics Conference*, pages 242–249, Aug. 2012. doi: 10.1109/EISIC.2012.12.

[90] S. K. Ernala, M. L. Birnbaum, K. A. Candan, A. F. Rizvi, W. A. Sterling, J. M. Kane, and M. De Choudhury. 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*, pages 1–16, 2019.

[91] H. J. Escalante, E. Villatoro-Tello, A. Juárez, M. Montes-y Gómez, and L. Villaseñor. Sexual predator detection in chats with chained classifiers. In *Proceedings of the 4th Workshop on*

Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 46–54, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W13-1607.

[92] M. A. Fauzi and P. Bours. Ensemble Method for Sexual Predators Identification in Online Chats. In *2020 8th International Workshop on Biometrics and Forensics (IWBF)*, pages 1–6, Apr. 2020. doi: 10.1109/IWBF49977.2020.9107945.

[93] C. Fellbaum. Wordnet. In *Theory and applications of ontology: computer applications*, pages 231–243. Springer, 2010.

[94] C. Fiesler and N. Proferes. "participant" perceptions of twitter research ethics. *Social Media+ Society*, 4(1):2056305118763366, 2018.

[95] C. Fiesler, J. R. Brubaker, A. Forte, S. Guha, N. McDonald, and M. Muller. Qualitative methods for CSCW. In *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing*. ACM, Nov. 2019. doi: 10.1145/3311957.3359428. URL https://doi.org/10.1145/3311957.3359428.

[96] R. J. Fisher. Social desirability bias and the validity of indirect questioning. *Journal of Consumer Research*, 20(2):303, Sept. 1993. doi: 10.1086/209351. URL https://doi.org/10.1086/209351.

[97] J. L. Fleiss and J. Cohen. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*, 33(3): 613–619, 1973.

[98] A. Forte, M. Dickard, R. Magee, and D. E. Agosto. What do teens ask their online social networks? In *Proceedings of the 17th ACM conference on Computer supported cooperative*

work & social computing. ACM, Feb. 2014. doi: 10.1145/2531602.2531723. URL https://doi.org/10.1145/2531602.2531723.

[99] R. Fournier and M. Danisch. Mining bipartite graphs to improve semantic pedophile activity detection. In *2014 IEEE Eighth International Conference on Research Challenges in Information Science (RCIS)*, pages 1–4, May 2014. doi: 10.1109/RCIS.2014.6861035.

[100] F. Frank and V. E. Xavier. Perverted-Justice.com - The largest and best anti-predator organization online. URL http://www.perverted-justice.com/.

[101] P. Friesen, L. Kearns, B. Redman, and A. L. Caplan. Rethinking the belmont report? *The American Journal of Bioethics*, 17(7):15–21, 2017.

[102] M. Gámez-Guadix, C. Almendros, E. Borrajo, and E. Calvete. Prevalence and association of sexting and online sexual victimization among spanish adults. *Sexuality Research and Social Policy*, 12(2):145–154, 2015.

[103] M. Gámez-Guadix, C. Almendros, E. Calvete, and P. D. Santisteban. Persuasion strategies and sexual solicitations and interactions in online sexual grooming of adolescents: Modeling direct and indirect pathways. *Journal of Adolescence*, 63:11–18, Feb. 2018. doi: 10.1016/j.adolescence.2017.12.002. URL https://doi.org/10.1016/j.adolescence.2017.12.002.

[104] H. Gardner and K. Davis. *The app generation: How today's youth navigate identity, intimacy, and imagination in a digital world*. 01 2013.

[105] N. Garg, L. Schiebinger, D. Jurafsky, and J. Zou. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16): E3635–E3644, 2018.

[106] G. D. P. R. (GDPR). Art. 20 gdpr – right to data portability | general data protection regulation (gdpr), 2021. URL https://gdpr-info.eu/art-20-gdpr/.

[107] A. K. Ghosh. Taking a more balanced approach to adolescent mobile safety. In *Proceedings of the 19th International Conference on Supporting Group Work*. ACM, Nov. 2016. doi: 10.1145/2957276.2997025. URL https://doi.org/10.1145/2957276.2997025.

[108] A. Ghosh Chowdhury, R. Sawhney, P. Mathur, D. Mahata, and R. Ratn Shah. Speak up, Fight Back! Detection of Social Media Disclosures of Sexual Harassment. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 136–146, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-3018. URL https://www.aclweb.org/anthology/N19-3018.

[109] L. Gillam and A. Vartapetiance. Quite Simple Approaches for Authorship Attribution, Intrinsic Plagiarism Detection and Sexual Predator Identification. In *LNCS*, Rome, Italy, Sept. 2012. URL http://www.uni-weimar.de/medien/webis/research/events/pan-12/pan12-web/.

[110] J. Goodwin and E. Tiderington. Building trauma-informed research competencies in social work education. *Social Work Education*, pages 1–14, 2020.

[111] A. Gorin and A. Stone. Recall biases and cognitive errors in retrospective self reports: A call for momentary assessments. 01 2001.

[112] A. M. G. Gualdo, S. C. Hunter, K. Durkin, P. Arnaiz, and J. J. Maquilón. The emotional impact of cyberbullying: Differences in perceptions and experiences as a function of role. *Computers & Education*, 82:228–235, 2015.

[113] M. Guillemin and L. Gillam. Ethics, reflexivity, and "ethically important moments" in research. *Qualitative inquiry*, 10(2):261–280, 2004.

[114] F. E. Gunawan, L. Ashianti, S. Candra, and B. Soewito. Detecting online child grooming conversation. In *2016 11th International Conference on Knowledge, Information and Creativity Support Systems (KICSS)*, pages 1–6, Nov. 2016. doi: 10.1109/KICSS.2016.7951413.

[115] B. Hallinan, J. R. Brubaker, and C. Fiesler. Unexpected expectations: Public reaction to the facebook emotional contagion study. *New Media & Society*, 22(6):1076–1094, Sept. 2019. doi: 10.1177/1461444819876944. URL https://doi.org/10.1177/1461444819876944.

[116] A. Hanna, E. Denton, A. Smart, and J. Smith-Loud. Towards a critical race methodology in algorithmic fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 501–512, 2020.

[117] E. Hargittai, J. Schultz, J. Palfrey, et al. Why parents help their children lie to facebook about age: Unintended consequences of the 'children's online privacy protection act'. *First Monday*, 2011.

[118] H. Hartikainen, A. Razi, and P. Wisniewski. Safe sexting: The advice and support adolescents receive from peers regarding online sexual risks. *PACM on Human Computer Interaction*, 2021.

[119] R. Hartson and P. Pyla. Chapter 22 - empirical ux evaluation: Ux goals, metrics, and targets. In R. Hartson and P. Pyla, editors, *The UX Book (Second Edition)*, pages 453–481. Morgan Kaufmann, Boston, second edition edition, 2019. ISBN 978-0-12-805342-3. doi: https://doi.org/10.1016/B978-0-12-805342-3.00022-9. URL https://www.sciencedirect.com/science/article/pii/B9780128053423000229.

[120] N. Hassan, A. Poudel, J. Hale, C. Hubacek, K. T. Huq, S. K. K. Santu, and S. I. Ahmed. Towards automated sexual violence report tracking. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 250–259, 2020.

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

[122] F. He, Y. Deng, and W. Li. Coronavirus disease 2019: What we know? *Journal of medical virology*, 92(7):719–725, 2020.

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

[124] J. M. G. Hidalgo and A. A. C. Díaz. Combining Predation Heuristics and Chat-Like Features in Sexual Predator Identification. page 6.

[125] L. C. Hillstrom. *The# metoo movement*. ABC-CLIO, 2018.

[126] M. Hind, S. Mehta, A. Mojsilovic, R. Nair, K. N. Ramamurthy, A. Olteanu, and K. R. Varshney. Increasing trust in ai services through supplier's declarations of conformity. *arXiv preprint arXiv:1808.07261*, 18:2813–2869, 2018.

[127] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman. Metrics for explainable ai: Challenges and prospects. *arXiv preprint arXiv:1812.04608*, 2018.

[128] A. Holzinger. Interactive machine learning for health informatics: when do we need the human-in-the-loop? *Brain Informatics*, 3(2):119–131, 2016.

[129] A. Z. Huq. Racial equity in algorithmic criminal justice. *Duke LJ*, 68:1043, 2018.

[130] M. Ibanez and R. Gazan. Virtual indicators of sex trafficking to identify potential victims in online advertisements. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 818–824. IEEE, 2016.

[131] M. Ibanez and R. Gazan. Detecting sex trafficking circuits in the U.S. through analysis of online escort advertisements. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 892–895, Aug. 2016. doi: 10.1109/ASONAM.2016.7752344.

[132] M. Ibanez and D. D. Suthers. Detection of domestic human trafficking indicators and movement trends using content available on open internet sources. In *2014 47th Hawaii international conference on system sciences*, pages 1556–1565. IEEE, 2014.

[133] M. Ibanez and D. D. Suthers. Detecting covert sex trafficking networks in virtual markets. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 876–879, Aug. 2016. doi: 10.1109/ASONAM.2016.7752340.

[134] G. Inches and F. Crestani. Overview of the international sexual predator identification competition at pan-2012. In *CLEF (Online working notes/labs/workshop)*, volume 30, 2012.

[135] G. Inches, M. Harvey, and F. Crestani. Finding participants in a chat: Authorship attribution for conversational documents. In *2013 International Conference on Social Computing*, pages 272–279. IEEE, 2013.

[136] F. Iqbal, B. C. M. Fung, M. Debbabi, R. Batool, and A. Marrington. Wordnet-Based Criminal Networks Mining for Cybercrime Investigation. *IEEE Access*, 7:22740–22755, 2019. ISSN 2169-3536. doi: 10.1109/ACCESS.2019.2891694.

[137] J. Isaak and M. J. Hanna. User data privacy: Facebook, cambridge analytica, and privacy protection. *Computer*, 51(8):56–59, 2018.

[138] S. Ishihara. A comparative study of likelihood ratio based forensic text comparison procedures: Multivariate kernel density with lexical features vs. word n-grams vs. character n-grams. In *2014 Fifth Cybercrime and Trustworthy Computing Conference*, pages 1–11. IEEE, 2014.

[139] A. Jaimes, D. Gatica-Perez, N. Sebe, and T. S. Huang. Guest editors' introduction: Human-centered computing–toward a human revolution. *Computer*, 40(5):30–34, 2007.

[140] H. Jia, P. J. Wisniewski, H. Xu, M. B. Rosson, and J. M. Carroll. Risk-taking As a Learning Process for Shaping Teen's Online Information Privacy Behaviors. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '15, pages 583–599, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-2922-4. doi: 10.1145/2675133.2675287. URL http://doi.acm.org/10.1145/2675133.2675287.

[141] E. S. Jo and T. Gebru. Lessons from archives: Strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 306–316, 2020.

[142] L. M. Jones, K. J. Mitchell, and D. Finkelhor. Trends in youth internet victimization: Findings from three youth internet safety surveys 2000–2010. *Journal of Adolescent Health*, 50(2):179–186, Feb. 2012. doi: 10.1016/j.jadohealth.2011.09.015. URL https://doi.org/10.1016/j.jadohealth.2011.09.015.

[143] M. Kang and S. Quine. Young people's concerns about sex: unsolicited questions to a teenage radio talkback programme over three years. *Sex Education*, 7(4):407–420, Nov. 2007. doi: 10.1080/14681810701636010. URL https://doi.org/10.1080/14681810701636010.

[144] S. Karlekar and M. Bansal. SafeCity: Understanding Diverse Forms of Sexual Harassment Personal Stories. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2805–2811, Brussels, Belgium, Oct. 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1303. URL https://www.aclweb.org/anthology/D18-1303.

[145] A. Khatua, E. Cambria, and A. Khatua. Sounds of Silence Breakers: Exploring Sexual Violence on Twitter. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 397–400, Aug. 2018. doi: 10.1109/ASONAM.2018.8508576. ISSN: 2473-9928.

[146] J. Kim, Y. J. Kim, M. Behzadi, and I. G. Harris. Analysis of Online Conversations to Detect Cyberpredators Using Recurrent Neural Networks. In *Proceedings for the First International Workshop on Social Threats in Online Conversations: Understanding and Management*, pages 15–20, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-39-9. URL https://www.aclweb.org/anthology/2020.stoc-1.3.

[147] J.-E. Kim, E. C. Weinstein, and R. L. Selman. Romantic relationship advice from anonymous online helpers. *Youth & Society*, 49(3):369–392, Aug. 2016. doi: 10.1177/0044118x15604849. URL https://doi.org/10.1177/0044118x15604849.

[148] S. Kim, A. Razi, G. Stringhini, P. Wisniewski, and M. De Choudhury. 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*, 2021.

[149] S. Kim, A. Razi, G. Stringhini, P. J. Wisniewski, and M. De Choudhury. A human-centered systematic literature review of cyberbullying detection algorithms. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–34, 2021.

[150] Y. Kim. Convolutional neural networks for sentence classification. In *EMNLP*, 2014.

[151] R. Kling and S. L. Star. Human centered systems in the perspective of organizational and social informatics. *SIGCAS Comput. Soc.*, 28(1):22–29, Mar. 1998. ISSN 0095-2737. doi: 10.1145/277351.277356. URL https://doi.org/10.1145/277351.277356.

[152] E. D. Klonsky and C. R. Glenn. 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):215–219, 2009.

[153] S. Kobayashi. Contextual augmentation: Data augmentation by words with paradigmatic relations. *arXiv preprint arXiv:1805.06201*, 2018.

[154] P. K. Kohler, L. E. Manhart, and W. E. Lafferty. Abstinence-only and comprehensive sex education and the initiation of sexual activity and teen pregnancy. *Journal of Adolescent Health*, 42(4):344–351, Apr. 2008. doi: 10.1016/j.jadohealth.2007.08.026. URL https://doi.org/10.1016/j.jadohealth.2007.08.026.

[155] A. Kontostathis. ChatCoder: Toward the Tracking and Categorization of Internet Predators. In *Proc. Text Mining Workshop 2009 Held in Conjunction with the Ninth Siam International Conference on Data Mining (sdm 2009). Sparks, Nv. May 2009.*, 2009.

[156] A. Kontostathis, L. Edwards, J. Bayzick, A. Leatherman, and K. Moore. Comparison of rule-based to human analysis of chat logs. *communication theory*, 8(2), 2009.

[157] P. Korenis and S. B. Billick. Forensic implications: Adolescent sexting and cyberbullying. *Psychiatric Quarterly*, 85(1):97–101, Oct. 2013. doi: 10.1007/s11126-013-9277-z. URL https://doi.org/10.1007/s11126-013-9277-z.

[158] P. Kostakos, L. Špráchalová, A. Pandya, M. Aboeleinen, and M. Oussalah. Covert Online Ethnography and Machine Learning for Detecting Individuals at Risk of Being Drawn into

Online Sex Work. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 1096–1099, Aug. 2018. doi: 10.1109/ASONAM.2018.8508276. ISSN: 2473-991X.

[159] M. A. Krieger. Unpacking "sexting": A systematic review of nonconsensual sexting in legal, educational, and psychological literatures. *Trauma, Violence, & Abuse*, 18(5):593–601, 2017.

[160] S. Krig. Ground truth data, content, metrics, and analysis. In *Computer Vision Metrics*, pages 247–271. Springer, 2016.

[161] K. Kroenke and R. L. Spitzer. The phq-9: a new depression diagnostic and severity measure, 2002.

[162] K. Krol, M. Moroz, and M. A. Sasse. Don't work. can't work? why it's time to rethink security warnings. In *Intl Conf on Risks and Security of Internet and Systems (CRiSIS)*, pages 1–8, 2012. doi: 10.1109/CRISIS.2012.6378951.

[163] C. Laorden, P. Galán-García, I. Santos, B. Sanz, J. Gomez Hidalgo, and P. Bringas. Negobot: A Conversational Agent Based on Game Theory for the Detection of Paedophile Behaviour. In *Advances in Intelligent Systems and Computing*, volume 189, pages 261–270. Jan. 2013. doi: 10.1007/978-3-642-33018-6_27.

[164] M. Latapy, C. Magnien, and R. Fournier. Quantifying paedophile queries in a large P2P system. In *2011 Proceedings IEEE INFOCOM*, pages 401–405, Apr. 2011. doi: 10.1109/INFCOM.2011.5935191. ISSN: 0743-166X.

[165] A. Lenhart. Teens, social media & technology overview 2015. 2015.

[166] A. Lenhart, M. Anderson, and A. Smith. Teens, Technology and Romantic Relationships

| Pew Research Center, Oct. 2015. URL http://www.pewinternet.org/2015/10/01/teens-technology-and-romantic-relationships/.

[167] A. Lenhart, A. Smith, and M. L. Anderson. Teens, technology and romantic relationships. 2015.

[168] L. Li, O. Simek, A. Lai, M. Daggett, C. K. Dagli, and C. Jones. Detection and Characterization of Human Trafficking Networks Using Unsupervised Scalable Text Template Matching. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 3111–3120, Dec. 2018. doi: 10.1109/BigData.2018.8622189.

[169] R. J. Light. Measures of response agreement for qualitative data: some generalizations and alternatives. *Psychological bulletin*, 76(5):365, 1971.

[170] H. Lim, D. G. Andersen, and M. Kaminsky. 3lc: Lightweight and effective traffic compression for distributed machine learning. *Proceedings of Machine Learning and Systems*, 1: 53–64, 2019.

[171] Y. Liu, Q. Li, X. Liu, Q. Zhang, and L. Si. Sexual Harassment Story Classification and Key Information Identification. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19, pages 2385–2388, Beijing, China, Nov. 2019. Association for Computing Machinery. ISBN 978-1-4503-6976-3. doi: 10.1145/3357384.3358146. URL http://doi.org/10.1145/3357384.3358146.

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

[173] N. Lorenzo-Dus, A. Kinzel, and M. Di Cristofaro. The communicative modus operandi of

online child sexual groomers: Recurring patterns in their language use. *Journal of Pragmatics*, 155:15–27, 2020. ISSN 0378-2166. doi: https://doi.org/10.1016/j.pragma.2019.09.010. URL https://www.sciencedirect.com/science/article/pii/S0378216619306162.

[174] N. Lorenzo-Dus, A. Kinzel, and M. Di Cristofaro. The communicative modus operandi of online child sexual groomers: Recurring patterns in their language use. *Journal of Pragmatics*, 155:15–27, Jan. 2020. ISSN 0378-2166. doi: 10.1016/j.pragma.2019.09.010. URL http://www.sciencedirect.com/science/article/pii/S0378216619306162.

[175] A. P. López-Monroy, F. A. González, M. Montes, H. J. Escalante, and T. Solorio. Early Text Classification Using Multi-Resolution Concept Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1216–1225, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1110. URL https://www.aclweb.org/anthology/N18-1110.

[176] X. Ma, J. Hancock, and M. Naaman. Anonymity, intimacy and self-disclosure in social media. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, May 2016. doi: 10.1145/2858036.2858414. URL https://doi.org/10.1145/2858036.2858414.

[177] K. MacFarlane and V. Holmes. Agent-Mediated Information Exchange: Child Safety Online. In *2009 International Conference on Management and Service Science*, pages 1–5, Sept. 2009. doi: 10.1109/ICMSS.2009.5302027. ISSN: null.

[178] A. Majchrzak, S. Faraj, G. C. Kane, and B. Azad. The contradictory influence of social

media affordances on online communal knowledge sharing. *Journal of Computer-Mediated Communication*, 19(1):38–55, 2013.

[179] C. D. Manning, M. Surdeanu, J. Bauer, J. R. Finkel, S. Bethard, and D. McClosky. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60, 2014.

[180] A. E. Marwick and D. Boyd. Networked privacy: How teenagers negotiate context in social media. *New Media & Society*, 16:1051 – 1067, 2014.

[181] M. Massimi, W. Moncur, W. Odom, R. Banks, and D. Kirk. Memento mori: technology design for the end of life. In *CHI'12 Extended Abstracts on Human Factors in Computing Systems*, pages 2759–2762. 2012.

[182] T. Matthews, K. O'Leary, A. Turner, M. Sleeper, J. P. Woelfer, M. Shelton, C. Manthorne, E. F. Churchill, and S. Consolvo. Stories from survivors. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, May 2017. doi: 10.1145/3025453.3025875. URL https://doi.org/10.1145/3025453.3025875.

[183] N. McDonald, K. Badillo-Urquiola, M. G. Ames, N. Dell, E. Keneski, M. Sleeper, and P. J. Wisniewski. Privacy and power: Acknowledging the importance of privacy research and design for vulnerable populations. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–8, 2020.

[184] I. McGhee, J. Bayzick, A. Kontostathis, L. Edwards, A. McBride, and E. Jakubowski. Learning to Identify Internet Sexual Predation. *International Journal of Electronic Commerce*, 15(3):103–122, Apr. 2011. ISSN 1086-4415. doi: 10.2753/JEC1086-4415150305. URL https://doi.org/10.2753/JEC1086-4415150305.

[185] B. C. McHugh, P. Wisniewski, M. B. Rosson, and J. M. Carroll. When social media traumatizes teens. *Internet Research*, 28(5):1169–1188, Oct. 2018. doi: 10.1108/intr-02-2017-0077. URL https://doi.org/10.1108/intr-02-2017-0077.

[186] S. K. Mckenzie, C. Li, G. Jenkin, and S. Collings. Ethical considerations in sensitive suicide research reliant on non-clinical researchers. *Research ethics*, 13(3-4):173–183, 2017.

[187] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*, 2019.

[188] M. W. R. Miah, J. Yearwood, and S. Kulkarni. Detection of child exploiting chats from a mixed chat dataset as a text classification task. In *Proceedings of the Australasian Language Technology Association Workshop 2011*, pages 157–165, Canberra, Australia, Dec. 2011. URL https://www.aclweb.org/anthology/U11-1020.

[189] D. Michalopoulos and I. Mavridis. Utilizing document classification for grooming attack recognition. In *2011 IEEE Symposium on Computers and Communications (ISCC)*, pages 864–869, June 2011. doi: 10.1109/ISCC.2011.5983950.

[190] D. Michalopoulos, E. Papadopoulos, and I. Mavridis. Artemis: Protection from Sexual Exploitation Attacks via SMS. In *2012 16th Panhellenic Conference on Informatics*, pages 19–24, Oct. 2012. doi: 10.1109/PCi.2012.46.

[191] P. Mishra, H. Yannakoudakis, and E. Shutova. Tackling online abuse: A survey of automated abuse detection methods. *arXiv preprint arXiv:1908.06024*, 2019.

[192] K. Misra, H. Devarapalli, T. R. Ringenberg, and J. T. Rayz. Authorship Analysis of Online Predatory Conversations using Character Level Convolution Neural Networks. In *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, pages 623–628, Oct. 2019. doi: 10.1109/SMC.2019.8914323. ISSN: 1062-922X.

[193] K. Mitchell, L. Jones, D. Finkelhor, and J. Wolak. Trends in unwanted online experiences and sexting : Final report. 2014.

[194] K. J. Mitchell. Risk factors for and impact of online sexual solicitation of youth. *JAMA*, 285 (23):3011, June 2001. doi: 10.1001/jama.285.23.3011. URL https://doi.org/10.1001/jama.285.23.3011.

[195] K. J. Mitchell and L. M. Jones. Youth internet safety study (yiss): Methodology report. 2011.

[196] K. J. Mitchell, D. Finkelhor, and J. Wolak. Online requests for sexual pictures from youth: Risk factors and incident characteristics. *Journal of Adolescent Health*, 41(2):196–203, Aug. 2007. doi: 10.1016/j.jadohealth.2007.03.013. URL https://doi.org/10.1016/j.jadohealth.2007.03.013.

[197] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229, 2019.

[198] M. Mladenović, V. Ošmjanski, and S. V. Stanković. Cyber-aggression, cyberbullying, and cyber-grooming: A survey and research challenges. *ACM Computing Surveys (CSUR)*, 54 (1):1–42, 2021.

[199] M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of machine learning*. MIT press, 2018.

[200] S. Mohseni, N. Zarei, and E. D. Ragan. A multidisciplinary survey and framework for design and evaluation of explainable ai systems. *arXiv preprint arXiv:1811.11839*, 2018.

[201] W. Moncur. The emotional wellbeing of researchers: considerations for practice. In *Pro-*

ceedings of the SIGCHI conference on human factors in computing systems, pages 1883–1890, 2013.

[202] M. A. Moreno, N. Goniu, P. S. Moreno, and D. Diekema. Ethics of social media research: common concerns and practical considerations. *Cyberpsychology, behavior, and social networking*, 16(9):708–713, 2013.

[203] C. Mori, J. E. Cooke, J. R. Temple, A. Ly, Y. Lu, N. Anderson, C. Rash, and S. Madigan. The prevalence of sexting behaviors among emerging adults: A meta-analysis. *Archives of sexual behavior*, pages 1–17, 2020.

[204] I. Nahum-Shani, S. N. Smith, B. J. Spring, L. M. Collins, K. Witkiewitz, A. Tewari, and S. A. Murphy. Just-in-time adaptive interventions (jitais) in mobile health: key components and design principles for ongoing health behavior support. *Annals of Behavioral Medicine*, 52(6):446–462, 2018.

[205] J. A. Naslund, K. A. Aschbrenner, L. A. Marsch, and S. J. Bartels. The future of mental health care: peer-to-peer support and social media. *Epidemiology and Psychiatric Sciences*, 25(2):113–122, Jan. 2016. doi: 10.1017/s2045796015001067. URL `https://doi.org/10.1017/s2045796015001067`.

[206] C. H. Ngejane, G. Mabuza-Hocquet, J. H. P. Eloff, and S. Lefophane. Mitigating Online Sexual Grooming Cybercrime on Social Media Using Machine Learning: A Desktop Survey. In *2018 International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD)*, pages 1–6, Aug. 2018. doi: 10.1109/ICABCD.2018.8465413.

[207] G. H. Nguyen, A. Bouzerdoum, and S. L. Phung. Learning pattern classification tasks with imbalanced data sets. *Pattern recognition*, pages 193–208, 2009.

[208] L. N. Olson, J. L. Daggs, B. L. Ellevold, and T. K. K. Rogers. Entrapping the innocent: Toward a theory of child sexual predators' luring communication. *Communication Theory*, 17(3):231–251, 2007. doi: https://doi.org/10.1111/j.1468-2885.2007.00294.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1468-2885.2007.00294.x.

[209] C. O'neil. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown, 2016.

[210] E. Ortiz-Ospina. The rise of social media. URL https://ourworldindata.org/rise-of-social-media.

[211] R. O'Connell. A typology of child cybersexploitation and online grooming practices. *Cyberspace Research Unit, University of Central Lancashire*, 2003.

[212] J. Palmer, D. Fam, T. Smith, and S. Kilham. Ethics in fieldwork: Reflections on the unexpected. *Qualitative Report*, 19(28):1–13, 2014.

[213] A. Panchenko, R. Beaufort, H. Naets, and C. Fairon. Towards Detection of Child Sexual Abuse Media: Categorization of the Associated Filenames. In P. Serdyukov, P. Braslavski, S. O. Kuznetsov, J. Kamps, S. Rüger, E. Agichtein, I. Segalovich, and E. Yilmaz, editors, *Advances in Information Retrieval*, Lecture Notes in Computer Science, pages 776–779. Springer Berlin Heidelberg, 2013. ISBN 978-3-642-36973-5.

[214] S. J. Pandey, I. Klapaftis, and S. Manandhar. Detecting Predatory Behaviour from Online Textual Chats. In A. Dziech and A. Czyżewski, editors, *Multimedia Communications, Services and Security*, Communications in Computer and Information Science, pages 270–281, Berlin, Heidelberg, 2012. Springer. ISBN 978-3-642-30721-8. doi: 10.1007/978-3-642-30721-8_27.

[215] J. Pater and E. Mynatt. Defining digital self-harm. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 1501–1513, 2017.

[216] J. Pater, C. Fiesler, and M. Zimmer. No humans here: Ethical speculation on public data, unintended consequences, and the limits of institutional review. *Proc. ACM Hum.-Comput. Interact.*, 6(GROUP), jan 2022. doi: 10.1145/3492857. URL `https://doi.org/10.1145/3492857`.

[217] C. Peersman, F. Vaassen, and V. V. Asch. Conversation Level Constraints on Pedophile Detection in Chat Rooms. page 13, 2012.

[218] N. Pendar. Toward Spotting the Pedophile Telling victim from predator in text chats. In *International Conference on Semantic Computing (ICSC 2007)*, pages 235–241, Sept. 2007. doi: 10.1109/ICSC.2007.32.

[219] L. Penna, A. Clark, and G. Mohay. A Framework for Improved Adolescent and Child Safety in MMOs. In *2010 International Conference on Advances in Social Networks Analysis and Mining*, pages 33–40, Aug. 2010. doi: 10.1109/ASONAM.2010.66.

[220] J. W. Pennebaker, M. E. Francis, and R. J. Booth. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001, 2001.

[221] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014. URL `http://www.aclweb.org/anthology/D14-1162`.

[222] R. M. Perhac Jr. Defining risk: Normative considerations. *Human and Ecological Risk Assessment*, 2(2):381–392, 1996.

[223] A. T. Pinter, P. J. Wisniewski, H. Xu, M. B. Rosson, and J. M. Caroll. Adolescent online safety. In *Proceedings of the 2017 Conference on Interaction Design and Children*. ACM, June 2017. doi: 10.1145/3078072.3079722. URL https://doi.org/10.1145/3078072.3079722.

[224] N. Potha, M. Maragoudakis, and D. Lyras. A biology-inspired, data mining framework for extracting patterns in sexual cyberbullying data. *Knowledge-Based Systems*, 96:134–155, Mar. 2016. ISSN 0950-7051. doi: 10.1016/j.knosys.2015.12.021. URL http://www.sciencedirect.com/science/article/pii/S0950705115005031.

[225] H. Pranoto, F. E. Gunawan, and B. Soewito. Logistic models for classifying online grooming conversation. *Procedia Computer Science*, 59:357–365, 2015.

[226] A. Rai. Explainable ai: From black box to glass box. *Journal of the Academy of Marketing Science*, 48(1):137–141, 2020.

[227] G. Ramos, J. Suh, S. Ghorashi, C. Meek, R. Banks, S. Amershi, R. Fiebrink, A. Smith-Renner, and G. Bansal. Emerging perspectives in human-centered machine learning. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–8, 2019.

[228] F. Rangel, P. Rosso, M. Koppel, E. Stamatatos, and G. Inches. Overview of the author profiling task at pan 2013. In *CLEF Conference on Multilingual and Multimodal Information Access Evaluation*, pages 352–365. CELCT, 2013.

[229] A. Razi, S. Kim, M. D. Choudhury, and P. Wisniewski. Ethical considerations for adolescent online risk detection ai systems. In *Good Systems: Ethical AI for CSCW (The 22nd ACM Conference on Computer-Supported Cooperative Work and Social Computing)*, 2019.

[230] A. Razi, Z. Agha, N. Chatlani, and P. Wisniewski. Privacy challenges for adolescents as

a vulnerable population. In *Networked Privacy Workshop of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020.

[231] A. Razi, K. Badillo-Urquiola, and P. J. Wisniewski. 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, pages 1–13, Honolulu, HI, USA, Apr. 2020. Association for Computing Machinery. ISBN 978-1-4503-6708-0. doi: 10.1145/3313831.3376400. URL http://doi.org/10.1145/3313831.3376400.

[232] A. Razi, S. Kim, A. Alsoubai, X. Caddle, S. Ali, M. D. Choudhury, P. Wisniewski, et al. Teens at the margin: Artificially intelligent technology for promoting adolescent online safety. In *ACM Conference on Human Factors in Computing Systems (CHI 2021)/Artificially Intelligent Technology for the Margins: A Multidisciplinary Design Agenda Workshop*, 2021.

[233] A. Razi, S. Kim, A. Soubai, G. Stringhini, T. Solorio, M. De Choudhury, and P. Wisniewski. A human-centered systematic literature review of the computational approaches for online sexual risk detection. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW2), Oct. 2021. doi: 10.1145/3479609. URL https://doi.org/10.1145/3479609.

[234] A. Razi, A. AlSoubai, S. Kim, N. Naher, S. Ali, G. Stringhini, M. De Choudhury, and P. J. Wisniewski. Instagram data donation: A case study on collecting ecologically valid social media data for the purpose of adolescent online risk detection. 2022.

[235] L. Reed, M. Boyer, H. Meskunas, R. Tolman, and L. Ward. How do adolescents experience sexting in dating relationships? motivations to sext and responses to sexting requests from dating partners. *Children & Youth Services Rev*, 109, 2020.

[236] E. Reeves. A synthesis of the literature on trauma-informed care. *Issues in mental health nursing*, 36(9):698–709, 2015.

[237] T. Ringenberg, K. Misra, K. C. Seigfried-Spellar, and J. Taylor Rayz. Exploring automatic identification of fantasy-driven and contact-driven sexual solicitors. In *2019 Third IEEE International Conference on Robotic Computing (IRC)*, pages 532–537, 2019. doi: 10.1109/IRC.2019.00110.

[238] T. R. Ringenberg, K. Misra, and J. T. Rayz. Not So Cute but Fuzzy: Estimating Risk of Sexual Predation in Online Conversations. In *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, pages 2946–2951, Oct. 2019. doi: 10.1109/SMC.2019.8914528. ISSN: 1062-922X.

[239] J. Ringrose, L. Harvey, R. Gill, and S. Livingstone. Teen girls, sexual double standards and 'sexting': Gendered value in digital image exchange. *Feminist Theory*, 14(3):305–323, Nov. 2013. doi: 10.1177/1464700113499853. URL https://doi.org/10.1177/1464700113499853.

[240] A. Romano. A new law intended to curb sex trafficking threatens the future of the internet as we know it. URL https://www.vox.com/culture/2018/4/13/17172762/fosta-sesta-backpage-230-internet-freedom.

[241] H. Rosa, D. Matos, R. Ribeiro, L. Coheur, and J. P. Carvalho. A "deeper" look at detecting cyberbullying in social networks. In *2018 international joint conference on neural networks (IJCNN)*, pages 1–8. IEEE, 2018.

[242] M. Rost, L. Barkhuus, H. Cramer, and B. Brown. Representation and communication: challenges in interpreting large social media datasets. page 6, 2013.

[243] T. Roy, J. McClendon, and L. Hodges. Analyzing Abusive Text Messages to Detect Digital

Dating Abuse. In *2018 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 284–293, June 2018. doi: 10.1109/ICHI.2018.00039.

[244] M. Rutgaizer, Y. Shavitt, O. Vertman, and N. Zilberman. Detecting Pedophile Activity in BitTorrent Networks. In N. Taft and F. Ricciato, editors, *Passive and Active Measurement*, Lecture Notes in Computer Science, pages 106–115, Berlin, Heidelberg, 2012. Springer. ISBN 978-3-642-28537-0. doi: 10.1007/978-3-642-28537-0_11.

[245] D. Sacha, M. Sedlmair, L. Zhang, J. A. Lee, D. Weiskopf, S. North, and D. Keim. Human-centered machine learning through interactive visualization. In *ESANN 2016: 24th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning Bruges, Belgium April 27-28-29, 2016 Proceedings*, pages 641–646, Bruges, Belgium, Aug. 2016. ESANN. ISBN 978-2-87587-026-1. URL `https://www.elen.ucl.ac.be/Proceedings/esann/esannpdf/es2016-166.pdf`.

[246] D. Saxena, K. Badillo-Urquiola, P. J. Wisniewski, and S. Guha. A human-centered review of algorithms used within the us child welfare system. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2020.

[247] B. W. Scharlott and W. G. Christ. Overcoming relationship-initiation barriers: The impact of a computer-dating system on sex role, shyness, and appearance inhibitions. *Computers in Human Behavior*, 11(2):191–204, 1995. doi: 10.1016/0747-5632(94)00028-g. URL `https://doi.org/10.1016/0747-5632(94)00028-g`.

[248] S. M. Schueller, M. Neary, J. Lai, and D. A. Epstein. Understanding people's use of and perspectives on mood-tracking apps: Interview study. *JMIR mental health*, 8(8):e29368, 2021.

[249] H. Schütze, C. D. Manning, and P. Raghavan. *Introduction to information retrieval*, volume 39. Cambridge University Press Cambridge, 2008.

[250] K. C. Seigfried-Spellar, M. K. Rogers, J. T. Rayz, S.-F. Yang, K. Misra, and T. Ringenberg. Chat Analysis Triage Tool: Differentiating contact-driven vs. fantasy-driven child sex offenders. *Forensic Science International*, Feb. 2019. ISSN 0379-0738. doi: 10.1016/j.forsciint.2019.02.028. URL http://www.sciencedirect.com/science/article/pii/S0379073818304420.

[251] S. Shahrokh Esfahani, M. J. Cafarella, M. Baran Pouyan, G. DeAngelo, E. Eneva, and A. E. Fano. Context-specific Language Modeling for Human Trafficking Detection from Online Advertisements. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1180–1184, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1114. URL https://www.aclweb.org/anthology/P19-1114.

[252] J. D. Shapka and R. Maghsoudi. Examining the validity and reliability of the cyber-aggression and cyber-victimization scale. *Computers in Human Behavior*, 69:10–17, Apr. 2017. ISSN 0747-5632. doi: 10.1016/j.chb.2016.12.015. URL http://www.sciencedirect.com/science/article/pii/S0747563216308354.

[253] Z. Shechtman, D. L. Vogel, H. A. Strass, and P. J. Heath. Stigma in help-seeking: the case of adolescents. *British Journal of Guidance & Counselling*, 46(1):104–119, 2018.

[254] D. R. Silva, A. Philpot, A. Sundararajan, N. M. Bryan, and E. Hovy. Data integration from open internet sources and network detection to combat underage sex trafficking. In *Proceedings of the 15th Annual International Conference on Digital Government Research*, pages 86–90, 2014.

[255] H. Sklenarova, A. Schulz, P. Schuhmann, M. Osterheider, and J. Neutze. Online sexual solicitation by adults and peers – results from a population based german sample.

Child Abuse & Neglect, 76:225–236, Feb. 2018. doi: 10.1016/j.chiabu.2017.11.005. URL https://doi.org/10.1016/j.chiabu.2017.11.005.

[256] S. Smirnova, S. Livingstone, and M. Stoilova. Understanding of user needs and problems: a rapid evidence review of age assurance and parental controls. 2021.

[257] T. Solorio, M. Shafaei, C. Smailis, I. Augenstein, M. Mitchell, I. Stapf, and I. Kakadiaris. White paper-creating a repository of objectionable online content: Addressing undesirable biases and ethical considerations.

[258] B. H. Spitzberg and G. Hoobler. Cyberstalking and the technologies of interpersonal terrorism. *New media & society*, 4(1):71–92, 2002.

[259] Stephanie. Cohen's kappa statistic. https://www.statisticshowto.datasciencecentral.com/cohens-kappa-statistic/, 2014. [Statistics How To. Retrieved June 27, 2019].

[260] S. Subramani, H. Wang, M. R. Islam, A. Ulhaq, and M. O'Connor. Child Abuse and Domestic Abuse: Content and Feature Analysis from Social Media Disclosures. In J. Wang, G. Cong, J. Chen, and J. Qi, editors, *Databases Theory and Applications*, Lecture Notes in Computer Science, pages 174–185, Cham, 2018. Springer International Publishing. ISBN 978-3-319-92013-9. doi: 10.1007/978-3-319-92013-9_14.

[261] H. Suresh, S. R. Gomez, K. K. Nam, and A. Satyanarayan. Beyond expertise and roles: A framework to characterize the stakeholders of interpretable machine learning and their needs. *arXiv preprint arXiv:2101.09824*, 2021.

[262] A. Suvarna, G. Bhalla, S. Kumar, and A. Bhardwaj. Identifying Victim Blaming Language in Discussions about Sexual Assaults on Twitter. In *International Conference on Social Media and Society*, SMSociety'20, pages 156–163, Toronto, ON, Canada, July 2020. Associa-

tion for Computing Machinery. ISBN 978-1-4503-7688-4. doi: 10.1145/3400806.3400825. URL http://doi.org/10.1145/3400806.3400825.

[263] L. K. Suzuki and J. P. Calzo. The search for peer advice in cyberspace: An examination of online teen bulletin boards about health and sexuality. *Journal of Applied Developmental Psychology*, 25(6):685–698, Nov. 2004. doi: 10.1016/j.appdev.2004.09.002. URL https://doi.org/10.1016/j.appdev.2004.09.002.

[264] M. U. Tariq, A. K. Ghosh, K. Badillo-Urquiola, A. Jha, S. Koppal, and P. J. Wisniewski. Designing Light Filters to Detect Skin Using a Low-powered Sensor. In *SoutheastCon 2018*, pages 1–8, Apr. 2018. doi: 10.1109/SECON.2018.8479027.

[265] M. U. Tariq, A. Razi, K. Badillo-Urquiola, and P. Wisniewski. A review of the gaps and opportunities of nudity and skin detection algorithmic research for the purpose of combating adolescent sexting behaviors. In *Lecture Notes in Computer Science*, pages 90–108. Springer International Publishing, 2019. doi: 10.1007/978-3-030-22636-7_6. URL https://doi.org/10.1007/978-3-030-22636-7_6.

[266] M. U. Tariq, A. Razi, K. Badillo-Urquiola, and P. Wisniewski. A Review of the Gaps and Opportunities of Nudity and Skin Detection Algorithmic Research for the Purpose of Combating Adolescent Sexting Behaviors. In M. Kurosu, editor, *Human-Computer Interaction. Design Practice in Contemporary Societies*, Lecture Notes in Computer Science, pages 90–108, Cham, 2019. Springer International Publishing. ISBN 978-3-030-22636-7. doi: 10.1007/978-3-030-22636-7_6.

[267] Y. R. Tausczik and J. W. Pennebaker. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1): 24–54, 2010.

[268] T. B. Team. Teen text speak codes every parent should know. https://www.bark.us/blog/teen-text-speak-codes-every-parent-should-know/, 2019.

[269] R. Tennant, L. Hiller, R. Fishwick, S. Platt, S. Joseph, S. Weich, J. Parkinson, J. Secker, and S. Stewart-Brown. The warwick-edinburgh mental well-being scale (wemwbs): development and uk validation. *Health and Quality of life Outcomes*, 5(1):1–13, 2007.

[270] D. L. Tolman and S. I. McClelland. Normative sexuality development in adolescence: A decade in review, 2000-2009. *Journal of Research on Adolescence*, 21(1):242–255, Feb. 2011. doi: 10.1111/j.1532-7795.2010.00726.x. URL https://doi.org/10.1111/j.1532-7795.2010.00726.x.

[271] F. Toriumi, T. Nakanishi, M. Tashiro, and K. Eguchi. Analysis of User Behavior on Private Chat System. In *2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, volume 3, pages 1–4, Dec. 2015. doi: 10.1109/WI-IAT.2015.49.

[272] A. Upadhyay, A. Chaudhari, a. S. Ghale, and S. S. Pawar. Detection and prevention measures for cyberbullying and online grooming. In *2017 International Conference on Inventive Systems and Control (ICISC)*, pages 1–4, Jan. 2017. doi: 10.1109/ICISC.2017.8068605.

[273] D. Van Bruwaene, Q. Huang, and D. Inkpen. A multi-platform dataset for detecting cyberbullying in social media. *Language Resources and Evaluation*, 54(4):851–874, 2020.

[274] T. F. Van de Mortel. Faking it: social desirability response bias in self-report research. *Australian Journal of Advanced Nursing, The*, 25(4):40–48, 2008.

[275] R. J. J. M. van den Eijnden, J. S. Lemmens, and P. M. Valkenburg. The Social Media Disorder Scale. *Computers in Human Behavior*, 61:478–487, Aug. 2016. ISSN 0747-

5632. doi: 10.1016/j.chb.2016.03.038. URL http://www.sciencedirect.com/science/article/pii/S0747563216302059.

[276] J. Van Ouytsel, E. Van Gool, K. Ponnet, and M. Walrave. Brief report: The association between adolescents' characteristics and engagement in sexting. *Journal of adolescence*, 37 (8):1387–1391, 2014.

[277] J. Van Ouytsel, M. Walrave, K. Ponnet, and W. Heirman. The association between adolescent sexting, psychosocial difficulties, and risk behavior: Integrative review. *The Journal of School Nursing*, 31(1):54–69, 2015.

[278] J. Van Ouytsel, N. M. Punyanunt-Carter, M. Walrave, and K. Ponnet. Sexting within young adults' dating and romantic relationships. *Current opinion in psychology*, 2020.

[279] A. Vartapetiance and L. Gillam. "Our Little Secret": pinpointing potential predators. *Secur Inform*, 3(1):3, Sept. 2014. ISSN 2190-8532. doi: 10.1186/s13388-014-0003-7. URL https://doi.org/10.1186/s13388-014-0003-7.

[280] J. W. Vaughan and H. Wallach. A human-centered agenda for intelligible machine learning. *Machines We Trust: Getting Along with Artificial Intelligence*, 2020.

[281] B. Vidgen, A. Harris, D. Nguyen, R. Tromble, S. Hale, and H. Margetts. Challenges and frontiers in abusive content detection. Association for Computational Linguistics, 2019.

[282] E. Villatoro-Tello, A. Juárez-González, H. J. Escalante, M. Montes-y Gómez, and L. Villaseñor-Pineda. A Two-step Approach for Effective Detection of Misbehaving Users in Chats. page 12.

[283] J. Vincett. Researcher self-care in organizational ethnography: Lessons from overcoming compassion fatigue. *Journal of Organizational Ethnography*, 2018.

[284] J. Vitak and J. Kim. " you can't block people offline" examining how facebook's affordances shape the disclosure process. In *In Proc 17th ACM CSCW*, pages 461–474, 2014.

[285] L. A. Voith, T. Hamler, M. W. Francis, H. Lee, and A. Korsch-Williams. Using a trauma-informed, socially just research framework with marginalized populations: practices and barriers to implementation. *Social Work Research*, 44(3):169–181, 2020.

[286] A. M. Walker, Y. Yao, C. Geeng, R. Hoyle, and P. Wisniewski. Moving beyond 'one size fits all'. *Interactions*, 26(6):34–39, Oct. 2019. doi: 10.1145/3358904. URL `https://doi.org/10.1145/3358904`.

[287] A. M. Walker, Y. Yao, C. Geeng, R. Hoyle, and P. Wisniewski. Moving beyond 'one size fits all': research considerations for working with vulnerable populations. *Interactions*, 26 (6):34–39, Oct. 2019. ISSN 10725520. doi: 10.1145/3358904. URL `http://dl.acm.org/citation.cfm?doid=3369909.3358904`.

[288] H. Wang, C. Cai, A. Philpot, M. Latonero, E. H. Hovy, and D. Metzler. Data Integration from Open Internet Sources to Combat Sex Trafficking of Minors. In *Proceedings of the 13th Annual International Conference on Digital Government Research*, dg.o '12, pages 246–252, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1403-9. doi: 10.1145/2307729.2307769. URL `http://doi.acm.org/10.1145/2307729.2307769`. event-place: College Park, Maryland, USA.

[289] J. Waycott, H. Davis, A. Thieme, S. Branham, J. Vines, and C. Munteanu. Ethical encounters in hci: Research in sensitive settings. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, pages 2369–2372, 2015.

[290] M. Webb, J. Burns, and P. Collin. Providing online support for young people with mental health difficulties: challenges and opportunities explored. *Early Intervention in Psychiatry*,

2(2):108–113, May 2008. doi: 10.1111/j.1751-7893.2008.00066.x. URL https://doi.org/10.1111/j.1751-7893.2008.00066.x.

[291] E. C. Weinstein and R. L. Selman. Digital stress: Adolescents' personal accounts. *New Media & Society*, 18(3):391–409, July 2014. doi: 10.1177/1461444814543989. URL https://doi.org/10.1177/1461444814543989.

[292] M. White. Receiving social support online: implications for health education. *Health Education Research*, 16(6):693–707, Dec. 2001. doi: 10.1093/her/16.6.693. URL https://doi.org/10.1093/her/16.6.693.

[293] H. Whittle, C. Hamilton-Giachritsis, A. Beech, and G. Collings. A review of online grooming: Characteristics and concerns. *Aggression and Violent Behavior*, 18(1):62–70, Jan. 2013. ISSN 1359-1789. doi: 10.1016/j.avb.2012.09.003. URL http://www.sciencedirect.com/science/article/pii/S1359178912001097.

[294] M. L. Williams, P. Burnap, and L. Sloan. Towards an ethical framework for publishing twitter data in social research: Taking into account users' views, online context and algorithmic estimation. *Sociology*, 51(6):1149–1168, 2017.

[295] E. Williamson, A. Gregory, H. Abrahams, N. Aghtaie, S.-J. Walker, and M. Hester. Secondary trauma: Emotional safety in sensitive research. *Journal of Academic Ethics*, 18(1): 55–70, 2020.

[296] T. Winograd. Shifting viewpoints: Artificial intelligence and human–computer interaction. *Artificial intelligence*, 170(18):1256–1258, 2006.

[297] P. Wisniewski, H. Jia, N. Wang, S. Zheng, H. Xu, M. B. Rosson, and J. M. Carroll. 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*, CHI '15, pages 4029–4038, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3145-6. doi: 10.1145/2702123.2702240. URL http://doi.acm.org/10.1145/2702123.2702240. event-place: Seoul, Republic of Korea.

[298] P. Wisniewski, H. Xu, M. B. Rosson, D. F. Perkins, and J. M. Carroll. 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, pages 3919–3930, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-3362-7. doi: 10.1145/2858036.2858317. URL http://doi.acm.org/10.1145/2858036.2858317. event-place: San Jose, California, USA.

[299] P. Wisniewski, A. K. Ghosh, H. Xu, M. B. Rosson, and J. M. Carroll. Parental control vs. teen self-regulation: Is there a middle ground for mobile online safety? In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 51–69, 2017.

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

[301] D. Y. Wohn. Volunteer moderators in twitch micro communities. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, May 2019. doi: 10.1145/3290605.3300390. URL https://doi.org/10.1145/3290605.3300390.

[302] P. Yan, L. Li, W. Chen, and D. Zeng. Quantum-Inspired Density Matrix Encoder for Sexual Harassment Personal Stories Classification. In *2019 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 218–220, July 2019. doi: 10.1109/ISI.2019.8823281. ISSN: null.

[303] M. L. Ybarra and K. J. Mitchell. How risky are social networking sites? a comparison of places online where youth sexual solicitation and harassment occurs. *Pediatrics*, 121(2): e350–e357, 2008.

[304] M. Young. Learning the art of helping: building blocks and techniques. 6. utg, 2017.

[305] P. Zambrano, J. Torres, L. Tello-Oquendo, R. Jácome, M. E. Benalcázar, R. Andrade, and W. Fuertes. Technical Mapping of the Grooming Anatomy Using Machine Learning Paradigms: An Information Security Approach. *IEEE Access*, 7:142129–142146, 2019. ISSN 2169-3536. doi: 10.1109/ACCESS.2019.2942805. Conference Name: IEEE Access.

[306] F. M. Zanzotto. Human-in-the-loop artificial intelligence. *Journal of Artificial Intelligence Research*, 64:243–252, 2019.

[307] J. M. Zhang, M. Harman, L. Ma, and Y. Liu. Machine learning testing: Survey, landscapes and horizons. *IEEE Transactions on Software Engineering*, 2020.

[308] Y. Zhang and X. Chen. Explainable Recommendation: A Survey and New Perspectives. *Foundations and Trends® in Information Retrieval*, 14(1):1–101, 2020. ISSN 1554-0669, 1554-0677. doi: 10.1561/1500000066. URL `http://arxiv.org/abs/1804.11192`. arXiv: 1804.11192.

[309] Z. Zuo, J. Li, P. Anderson, L. Yang, and N. Naik. Grooming Detection using Fuzzy-Rough Feature Selection and Text Classification. In *2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–8, July 2018. doi: 10.1109/FUZZ-IEEE.2018.8491591.