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

By

Ashwaq Alsoubai

Dissertation
Submitted to the Faculty of the
Graduate School of Vanderbilt University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY
in

Computer Science

May 10, 2024

Nashville, Tennessee

Approved:

Pamela Wisniewski, Ph.D.

Meiyi Ma, Ph.D.

Hiba Baroud, Ph.D.

Gianluca Stringhini, Ph.D.

This dissertation is dedicated with profound appreciation and love to my family, whose unwavering support and sacrifice have been the cornerstone of my journey. To my parents, Thawab and Sarah, whose lessons in strength, dedication, and self-motivation have been my guiding light, instilling in me the resilience to pursue my dreams. To my husband, Faraj, whose sacrifices and unyielding support have been my sanctuary, providing me with the strength and peace needed to navigate this challenging journey. And to my children, Mohammed and Abdullah, whose innocent smiles and boundless joy have been my refuge, reminding me of the beauty and simplicity of life amidst the complexities of academic pursuit. Your collective love and support have been the wind beneath my wings, propelling me forward and upward. This achievement is not just mine but a testament to the enduring strength and love of our family.

ACKNOWLEDGMENTS

This research is supported in part by the U.S. National Science Foundation under grants #IIP-2329976, #IIS-2333207, #CNS-1942610 and by the William T. Grant Foundation grant #187941. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the research sponsors. We would also like to thank all the participants who donated their data and contributed to the research.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction

1.1 Adolescents Digital Experiences and Risks

Most youths' social and developmental growth has been mediated through social media extensive usage (8). While the mediated access to such platforms enables them to experience important learning and communication skills, it also exposes them to a broader array of risks than before (207). Based on recent Pew Research findings (27), almost half of U.S. teens ages 13 to 17 (46%) reported experiencing at least one of six online harassment behaviors, with offensive name-calling and spreading of false rumors about them being the most commonly faced risks. In 2022, The annual report by Bark technologies (337) reported that 90.73% of teens came across nudity or sexual content on online platforms, 64.3% of them encountered or engaged with self-harm/suicidal content online, and 88.2% of them used online platforms to express or experience violent subject or thoughts.

In light of the prevalence of these online risks, the literature on adolescent online safety has taken a "risk-adverse" lens to primarily focus on protecting adolescents from exposure to possible threats (360). An overly generalized view of adolescent online risk experiences risk can impede prevention efforts, and intervention plans to protect teens from experiencing adverse life events that could negatively impact them into adulthood. For instance, legal scholars have generated a substantial body of work on laws to criminalize and appropriately respond to technology-facilitated sexual offenses (154; 82; 155). These works highlight how civil and criminal laws have fallen behind on how to respond effectively to online sexual risks, especially those involving minors, due to the lack of comprehensive understanding of these risks and the harms associated with them (82; 155). While these works make great strides in addressing the potential legal consequences, complexities, and pitfalls of online sexual risk behavior, they are not always grounded in the first-person online experiences of youth. Another example can be seen in the development of machine learning algorithms that predict and detect online sexually risky behavior, such as child pornography, sexual harassment, predation, or solicitations (373; 48; 139), which may cause unintentional harm, particularly to more vulnerable users, if these experiences are not appropriately contextualized.

Additionally, The research community has shown a great interest in understanding the boundaries between the offline and online behaviors of youth, mainly in the context of these risk experiences (106). In fact, offline and online risk experiences of youth were found to be driven by the same underlying factors related to the general propensity to experience risks (135). Findings from different studies have pointed to significant positive correlations between youth online and offline risk experiences such as between online sexual risks and self-harm (242; 353) or substance misuse (40), or between online harassment and self-harm (311) or substance misuse (275; 380). Therefore, in an increasingly digitized society, the lives of modern-day youth are ever-complex as they encounter new opportunities and risks offline, online, and at the intersection of these two converging realms. The risk prevention strategies that were developed so far to keep youth safe both offline and online need to do a better job of reflecting these complicated interactions and multi-faceted risk behaviors and experiences.

1.2 A Case for Human-Centered Machine Learning

Recently, Artificial Intelligence (AI)-based risk detection models have been presented as a potential solution to mitigate the online risks that youth encounter such as sexual risks, cyberbullying, and self-harm (13; 180). Such detection approaches should be harnessed to translate youths' behaviors, and societal and psychological needs into practical solutions. This will ensure digital equity, especially for socio-economically disadvantaged youth (289). Therefore, a sub-field of Computer Science called "Human-Centered Machine Learning" (HCML) has emerged to leverage human knowledge into the computing automatic approaches to address societal needs and enhance the practicality and applicability of these approaches (163). Human-centeredness leverages human insights and understanding to address the potential pitfalls and ethical considerations inherent in relying on a purely computational lens to solve human problems and to provide a better understanding of how these models will perform in real scenarios and the potential impact these computational solutions will have on end users (163). Moreover, Human-Centered Design (HCD) allows researchers to incorporate the perspectives of various stakeholders, including the survivors of risks, to construct a more efficient algorithm (163); in our context, leveraging this human-centered lens can mediate detecting, mitigating, and preventing online risks that youth encounter in a more nuanced way.

Upon reviewing the current machine learning risk detection approaches incorporating a human-centered lens, multiple gaps have been identified. Most of these detection models: 1) are built using public datasets that are not ecologically valid (i.e., not representative of youths' language) (290), 2) lack youths' risk perspectives when identifying ground truth annotations (179), and more importantly, 3) lack the comprehensive understanding of the societal and psychological context of risks to identify patterns of risks that could be used to prevent any occurrence of victimization before or when it happens (290). Human-centered real-time risk detection is crucial for youth to be able to provide them with in-time treatment resources, support, and interventions in effective ways (5). In fact, there is an increased interest in deploying human-centered risk detection within the SIGCHI and Computer-Supported Cooperative Work and Social Computing (CSCW) research communities, such as Razi et al. (290) in leveraging societal context to detect online sexual risks and Ali et al. who provided a multi-model approach to detect the risks such as harassment and sexual solicitation that youth encounter online (13). While these human-centered risk detection approaches provide accurate risk detection, there is still a need to count for the risk detection timing, which is a critical aspect to proactively protecting youth online, which warrants our sustained attention. In Chapter 3, we build upon this prior research by first taking a narrow-angle of risk to focus on analyzing youth public disclosures about their online sexual risk from an online peer support platform to holistically understand the nuances of youths' online sexual risks to operationalize this knowledge for building models. In Chapter 4, based on the results of our study one that youth sexual risks could laden by mental health indicators, we took a broader angle to take into account and investigate the multidimensional nature of the risk experiences of youth across online and offline contexts. Then in chapter 5, based on an identified appropriate youth-centered algorithmic framework to advance the real-time risk detection models. This process incorporates youths' societal, psychological, temporal, and linguistic-driven features that are indicative of risks. The following are associated research questions that will be examined in this dissertation:

- **RQ1-Literature Review:** *What are the trends, gaps, and opportunities in the current literature of computational approaches for real-time online 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?*

- **RQ2-Study 1:** *What insights can we gain regarding the online risk experiences of youth through their disclosures of sexual risk experiences when seeking peer support online?*

- **RQ3- Study 2:** *How does creating profiles of youth based on their self-reported online and offline risk behaviors inform about the multidimensionality of their lived risk experiences on social media?*

- **RQ4- Study 3:** *Can the insights gained from the prior two studies be built upon to create algorithms that accurately detect high-risk conversations in real-time?*

1.3 Dissertation Overview

In this dissertation, we direct our attention toward a better understanding of the risk experiences of adolescents and the development of algorithms that can detect these risks accurately and timely. To address these research questions, this dissertation conducted three studies that leveraged a mixed-methods approach: statistical analysis, qualitative analysis, and machine learning algorithms (topic modeling, sentiment analysis, and Natural Language Processing (NLP)). In Chapter 2, we describe a systematic literature review that we conducted to answer RQ1 which summarizes the trends, gaps, and opportunities within the existing real-time risk detection approaches for social media platforms. Then, based on the insights from reviewing the articles we presented our computational framework for human-centered, timely, and accurate online risk detection. In Chapter 3, we address RQ2 by conducting a content analysis of 45,955 posts by adolescents on an online peer support mental health forum to understand the nuances of online sexual risks of youth. In Chapter 4, we address RQ3 by presenting the profiles of youth based on their online and offline risk experiences, including online sexual risks, harassment, and offline risk behaviors including substance misuse and self-harm. In Chapter 5, we answer RQ4 by conducting experiments on an ecologically valid data set to build human-centered machine learning algorithms for various online risks detection in adolescents' private conversations, including sexual, harassment, and offline risk manifested in online spaces. Lastly, in Chapter 6, we provide an overview of the expected outcomes from this dissertation and a timeline for completion.

CHAPTER 2

Systemization of Knowledge (SoK): Creating a Research Agenda for Human-Centered Real-Time Risk Detection on Social Media Platforms

Citation: Ashwaq Alsoubai, Jinkyung Park, Afsaneh Razi, Sarvech Qadir, Gianluca Stringhini, Pamel J Wisniewski. Systemization of Knowledge (SoK): Creating a Research Agenda for Human-Centered Real-Time Risk Detection on Social Media Platforms. Proceedings of the ACM On Human Factors in Computing Systems (CHI 2024).

Accurate real-time risk identification is vital to protecting social media users from online harm, which has driven research towards advancements in machine learning (ML). While strides have been made regarding the computational facets of algorithms for "real-time" risk detection, such research has not yet evaluated these advancements through a human-centered lens. To this end, we conducted a systematic literature review of 53 peer-reviewed articles on real-time risk detection on social media. Real-time detection was mainly operationalized as "early" detection after-the-fact based on pre-defined chunks of data and evaluated based on standard performance metrics, such as timeliness. We identified several human-centered opportunities for advancing current algorithms, such as integrating human insight in feature selection, algorithms' improvement considering human behavior, and utilizing human evaluations. This work serves as a critical call-to-action for the HCI and ML communities to work together to protect social media users before, during, and after exposure to risks.

2.1 Introduction

Within the evolving field of Human-Centered Machine Learning (HCML), scholars have highlighted the need to keep machine learning algorithms grounded in human social and psychological needs (66), to minimize bias by being more inclusive to adequately represent the myriad of individuals' experiences, and to incorporate transparency and interpretability to understand the potential harms that could be caused to people by these algorithms (39; 145; 103). As such, there has been a shift in which scholars within the SIGCHI research communities have begun to apply a human-centered lens to synthesize and critique computational approaches for various forms of automated risk detection, including but not limited to online harassment, unwanted sexual solicitations, and mental health disclosures that occur via social media platforms (c.f., (290; 333; 179; 339; 10)). Indeed, social media has become a prominent part of people's lives that allows users to connect with others, share, create, and engage with various forms of digital content (9). In 2022, there were 3.96 billion social media users who spent hours a day using various social media platforms (e.g., Facebook, Twitter, Instagram) (32), demonstrating how social media is now an irrevocable and important part of our daily lives.

While social media can undeniably be beneficial, it can also facilitate digitally-mediated risks (e.g., amplifying misinformation, mental health challenges, and interpersonal violence (254; 132; 351; 309; 269)) that cannot be ignored. As a case in point, Meta (i.e., Facebook) recently faced significant criticism and legal scrutiny after the release of internal documents[1] suggested that the company took inadequate actions to mitigate risks related protecting youth from sexual exploitation, preventing human and drug trafficking, unfairly influencing political outcomes, and other harms to users and society-at-large (112). Yet, Meta is not alone regarding such criticism (278), nor have they or other social media platforms been inactive in addressing these mounting concerns. Legislators are also grappling with how to make social media companies accountable for the harms that occur on their platforms and have introduced several bills (e.g., Kids Online Safety Act (85) and Fight Online Sex Trafficking Act (172)) to proactively protect social media users from informational, mental health, and physical threats that have been at the forefront of news media reports, and consequently, machine learning (ML) and Human-Computer Interaction (HCI) research. These issues are hard to tackle because of the complexities inherent in human behavior that may result in risk misidentification.

Timely and accurate risk identification is necessary to effectively prevent harm and to provide a safe environment for all social media users (125). Therefore, recent advancements in ML and automated risk detection have grown beyond the traditional supervised learning paradigm to tackle more complex and dynamic problems, such as "real-time" risk detection (251). The real-time aspect of detection is important as identification must occur as early as possible to prevent the spread of the risk (e.g., fake news) or to mitigate harm as a result of it (e.g., mental health problems). Therefore, several recent works (296; 284) on computational machine-learning algorithms for real-time risk detection on social media have called for more research to advance the *technical aspects* of real-time risk detection to optimize performance. These works provide valuable insights about advancing the computational approaches for real-time risk detection; yet, there is still a need to evaluate real-time algorithms for detecting risks from a human-centered perspective (e.g., whether and how such algorithms can impact users in the real world) to ensure that they can effectively be leveraged as a first defense to prevent online harm, as opposed to contributing to it.

To apply such a human-centered perspective, we leveraged and extended Razi et al.'s (290) generalizable framework established for conducting human-centered systematic reviews of computational risk detection research. We apply this framework to the novel context of real-time risk detection within social media and augment it by adding new dimensions (e.g., input prioritization, timeliness) related to 'real-time' risk detection, which has been conceptualized and operationalized in the literature in multiple, and at times, conflicting ways. In this paper, we considered both computational and human-centered aspects of this literature to create

[1] https://www.wsj.com/articles/the-facebook-files-11631713039

a forward-thinking research agenda that advances our capacity to proactively protect social media users from online harm as it unfolds. As such, we set forth to answer the following high-level research questions:

- **RQ1:** *How has 'real-time' social media risk detection been defined and operationalized in the literature?*
- **RQ2:** *What are the state-of-the-art computational trends for real-time risk detection on social media?*
- **RQ3:** *Using a human-centered lens, what are the potential gaps and areas for future research for real-time risk detection on social media?*

To answer these questions, we systematically reviewed 53 peer-reviewed papers published between 2015 and 2023 that tackled core aspects of 'real-time' risk detection using social media data. We broadly considered all types of social media risks that may result in individual-level harm (e.g., mental health, sexual solicitations) or community-level harm (e.g., fake news, misinformation). We qualitatively coded the articles to answer our research questions. Overall, we found that real-time risk detection has been predominantly operationalized as early risk detection after-the-fact, but as early as possible (RQ1). For RQ2, the computational trends in prior studies included utilizing publicly available large-scale datasets, using commonly known machine learning features, improving deep learning-based approach, presenting the performance evaluation metrics mainly using pre-defined chunks of data. For RQ3, we identified gaps and opportunities for future research to advance these computational approaches based on a human-centered perspective. Opportunities included placing humans at the center of data collection and model evaluation endeavors, with the aim of comprehending their behaviors to serve as the foundation for the selection of features and the optimization of real-time risk detection models. To synthesize our findings, we created a cohesive framework to direct future research on creating efficient and human-centered real-time risk detection algorithms for social media. Our systemization of knowledge makes the following novel contributions to the HCI, HCML, and ML research communities:

- Formally defined and expanded the term "real-time" within the context of social media risk detection by incorporating a spectrum of detection mechanisms to detect the risk before, early, and to mitigate harm after-the-fact.
- Highlighted trends and best technical practices of the existing state-of-the-art computational approaches for real-time risk detection.
- Extended Razi et al. (290) and discovered potential gaps within the literature and provided agendas toward human-centered real-time risk detection that goes beyond the current state-of-the-art computational risk detection.

- Established a research agenda for advancing real-time computational risk detection to address both the computational challenges and human-centered gaps.

Next, we will synthesize the related work that motivated the need for this systematic literature review.

2.2 Background

In this section, we synthesize the current literature on trends within computational approaches for social media risk detection, timeliness in social media risk detection, and human-centered lens to review risk detection algorithms.

2.2.1 Trends within Computational Approaches for Social Media Risk Detection

Prior literature reviews of computational risk detection on social media focused predominantly on evaluating the data collection and prepossessing techniques, feature engineering process, algorithms, and common machine learning metrics for benchmarking performance (69; 274; 372; 140). A major number of these reviews have been centered around evaluating which machine learning algorithm performed best for detecting risks on social media. Traditional algorithms such as Logistic Regression, Support Vector Machines, Random Forest, and Naive Bayes have been extensively used in detecting risks in various social media platforms (69; 113; 140; 26; 377; 12). However, given the massive scale of social media data, these algorithms were found to not adapt well to the evolving patterns of the risks and struggled to handle large volumes of data (38). In recent years, there has been an increased interest in leveraging deep learning models, particularly suited for large-scale and complex datasets such as social media dataset (120; 220). For instance, Dowlagar and Mamidi (104) found that transformers with selective translation demonstrated promising results compared to other common neural network-based models. As such, prior research has pointed to the importance of capturing the intricate patterns and evolving trends of risks on social media, which could mainly be accomplished through leveraging dynamic, novel, and specialized techniques for annotating datasets, crafting features, or enhancing models (12). In fact, these techniques were shown to enhance risk detection capabilities and provide more reliable and effective risk management solutions in social media contexts. For instance, Yi and Zubiaga's (377) showed that novel models (i.e., MMCD (279) and XBully (78)) outperformed all pre-trained language models (e.g., BERT). Expanding beyond these works, in this review, we focused on identifying the trends within novel computational approaches, rather than off-the-shelf models, to detect the rapidly changing nature of risks within social media in real time.

2.2.2 Research on ‘Real-Time’ Risk Detection on Social Media

Research has called for current and future efforts on detecting and mitigating risks in social media to move towards building real-time risk detection systems (290; 179). One of the main efforts toward presenting and evaluating timely risk detection was introduced by the Early Risk Prediction on the Internet (eRisk) group, to examine methodologies and metrics related to early risk detection. Based on their yearly events, several reviews (209; 210; 211; 212; 265; 266) have been published to evaluate the timing risk detection for a myriad of issues (e.g., depression, self-harm, pathological gambling, and eating disorders) using social media data. Real-time solutions lie in performance-oriented evaluations; for example, detection time (101); yet, a remaining question is what the real-time aspects of these models are that set them apart from more traditional and cross-sectional approaches for computational risk detection.

Ample literature within computer science has focused on specifying the definition of “real-time” problem-solving (260; 189; 317; 101). Examples of common real-time definitions were “there is a strict time limit by which a system must have produced a response, regardless of the algorithm employed” (260) and “ability of the system to guarantee a response after a (domain defined) fixed time has elapsed (189).” These definitions carry flexibility, leaving room for varied interpretations based on application. Therefore, researchers have attempted to identify specific components of real-time systems (317; 101). For example, Shin et al (317) presented three main components of real-time systems, which were *time*, arguably the most important aspect of real-time systems, *task* that must be accomplished before the deadline, and *message or response* that should be received in a timely manner. Given these efforts of identifying real-time components, a more unified and comprehensive definition of real-time was still needed to provide clarity and precision. Therefore, Bruda and Akl (57) presented a formal and unified theory of real-time definition that was generalizable across domains. This theory consisted of two concepts centered on real-time systems: “computing with deadlines, and input data that arrive in a sequential manner or real-time”, which is the definition we adopted for our review.

2.2.3 Using a Human-Centered Lens to Review Real-Time Risk Detection Algorithms

The SIGCHI community has recently exhibited a growing interest in human-centered reviews, aiming to assess the effectiveness and impact of algorithms in real-world contexts (e.g., (290; 179; 308; 66)). Scholars have presented systematic reviews that underscore the importance of a human-centered approach to online risk detection, focusing on specific topics such as cyberbullying (179), sexual risks (290), child welfare system (308) or mental health (66). In the context of misinformation detection, Das et al. (93), reviewed NLP approaches for fact-checking from different human-centered strategies. They suggested guiding technology development for human use and practical adoption, and human-centered design practices early in model development. In another work, using a three-prong human-centeredness algorithm design framework, Kim et al. (179) analyzed cyberbullying detection approaches and found a lack of human-centeredness in defining cyberbullying, establishing ground truth in data annotation, assessing detection models' performance, speculating the uses and users of the models, including potential negative consequences. These prior reviews highlighted that human-centered approaches to evaluating risk detection algorithms are pivotal to ensure that the algorithms are designed to benefit those who are negatively affected by the risks the most (290), and how their involvement in research can lead to more practical, widely accessible solutions catering to individuals' diverse needs (179).

In this paper, we adopted the human-centered framework proposed by Razi et al. (290) for systematically evaluating real-time risk detection algorithms in terms of: 1) characteristics of the **dataset**, 2) pre-processing and **model development**, 3) **evaluation**, and 4) **applications and interventions** (As shown in Table 2.1). While our work leverages Razi et al.'s framework for how to conduct a human-centered review of computational risk detection research, the scope of our research differs. Razi et al.'s work focused solely on online sexual risks, similar to the other human-centered reviews of computational risk detection that focused on singular risk types (e.g., online harassment (179), and mental health (66)). In contrast, our review synthesizes computational approaches *across multiple risk types*. This approach allows us to consider these risk detection algorithms' broader implications and potential consequences in real-world settings, ensuring the development of more effective and socially responsible solutions. Most importantly, while these systematic and human-centered reviews (290; 179; 66) covered many aspects of computational risk detection, in general, they did not specifically focus on the *real-time* aspect of risk detection, which is the novel focus and contribution of this paper.

2.3 Methods

Below, we describe our systematic review of the literature and our qualitative synthesis of the articles in our dataset.

2.3.1 Systematic Literature Search Process

We selected five electronic databases (i.e., IEEE Xplore Digital Library, ACM Digital Library, ScienceDirect, Springer-link, and ACL Anthology) that ranged in computational and social science research approaches for the initial literature search to ensure broad coverage. We searched using combinations of keywords at the intersection of 1) social media (i.e., "social media", "Twitter', "Facebook", "Instagram", "YouTube"), 2) real-time detection (i.e., "real-time", "forecast", "early detection", "early prediction", "proactive prediction", or "proactive detection"), 3) online risks ("risk detection", "mental health", "cyberbullying", "sexual", "hate speech", "fake news", "incivility", "harassment", "abuse", or "spam"), and 4) computational approaches (“machine learning”, “artificial intelligence”, “deep learning”, or “algorithm”). We did not restrict the filter to a given date range. The initial search resulted in 2,212 papers, where 48% of the papers were from the ACM Digital Library. To confirm relevancy, we read through the papers’ titles, abstracts, methods, results, and conclusions based on the following inclusion criteria:

1. The paper was a peer-reviewed published work. Journal articles and conference proceedings were both included.

2. The paper should not be a purely theoretical analysis or summarize or evaluate existing studies (e.g., reviews).

3. The paper focused on social media risk detection. We used a wide angle of prevalent social media risks since this literature review focuses on real-time detection approaches, not online risks. Social media was selected due to the affordance of open-source data, which made these platforms a popular choice for researchers to apply risk detection approaches (321).

4. The paper presented a real-time approach using a computational or algorithmic approach such as Natural Language Processing (NLP) or Machine Learning (ML) that consists of both aspects of real-time models’ definition: 1) sequential input and 2) timeliness (57).

5. The paper provided a new computational approach or an enhancement of an existing approach, rather than only training or fine-tuning off-the-shelf computational models that are mainly designed for general use cases and may not provide the same level of precision when applied to specific domains like real-time risk detecting in social media.

We coded a paper as relevant if it met all the criteria above, which resulted in 45 papers. To identify additional relevant papers that were not yielded in our initial search, we cross-referenced the citations of the relevant papers. This cross-referencing resulted in 33 unique papers that were potentially relevant, of which 8 papers met our inclusion criteria. After one more iteration of this process, no additional relevant papers were identified, which suggests that we reached a saturation point. Therefore, our final search resulted in **53** relevant papers for our review.

2.3.2 Data Analysis Approach

To answer RQ1, we used a thematic analysis approach (50) to code papers related to how real-time risk detection was defined. To answer RQ2 and RQ3, we utilized the human-centered framework presented by Razi et al. (290) to review papers based on the 1) ecological validity of the dataset for detecting the risks, 2) investigating if the models are grounded in human theory, knowledge, and understanding, 3) performance of algorithms in terms of meeting end users' needs, 4) their outcomes when deployed in real-world settings. Razi's et al. (290) framework was created based on computational sexual risk detection; therefore, while coding the papers, we identified and added emerging codes that were not covered in this framework and suited our *real-time* detection for a generalized view of risks. We iteratively created new codes to suit the real-time aspect of the detection process. Codes were allowed to overlap for double-coding. Two coders labeled the same 10% of articles, and we calculated Cohen's Kappa IRR (83) to ensure the robustness of our coding process, which was 0.87. The researchers met to discuss the articles to resolve conflicts. Once a consensus was reached, the codes were updated. The remaining articles were then divided and coded by the two coders. The first author identified key themes by reviewing and conceptually grouping the final codes. The definitions of our codes and grounded codes that emerged from our data are shown in Table 2.1.

2.4 Results

We present the characteristics of our dataset, followed by our results organized by our over-arching research questions.

2.4.1 Defining and Operationalizing "Real-time" Risk Detection on Social Media

2.4.1.1 Types of Risks Detected

As illustrated by Figure 4.3, the research papers considered three main types of risk: fake news (55%), cyberbullying (30%), and mental illness (15%). There was a pronounced surge in articles on the real-time detection of fake news in the years 2020 and 2021. This time span coincided with the COVID-19 pandemic and the concurrent escalation in both rumors and fake news dissemination (148). This alignment potentially contributed to the escalated scholarly output during this period to combat all kinds of fake news, including rumors and misinformation. The historical significance of 2017 for the increase in the number of publications on real-time risk detection for mental health could be aligned with the launch of the ERISK workshop, [2] as discussed in our Background section. A common theme among these papers was the incorporation of the time-evolving aspect of the risk when building risk detection algorithms. For instance, cyberbullying implies a repetitive behavior over time (208), while mental illness symptoms, such as eating disorders and depression appear for a longer period of time to ensure correct diagnosis of mental illness (283; 375). Additionally, papers that addressed fake news and rumor detection presented the timing as the spread or propagation of such content throughout the network.

[2] https://erisk.irlab.org/2017/index.html

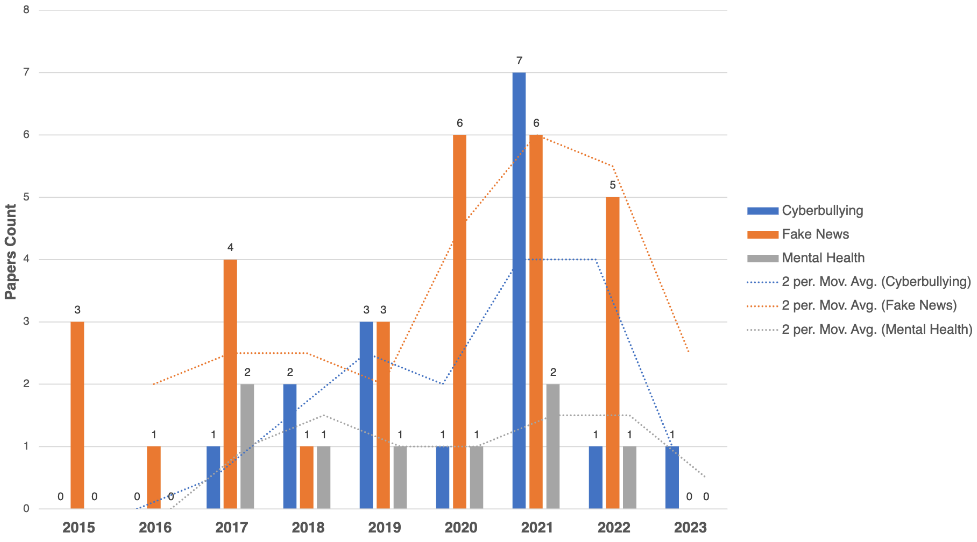


Figure 2.1: Number of Publications by Risk Type Over Time.

2.4.1.2 Definitions and Operationalization of ‘Real-Time’ Risk Detection

For RQ1, we first set out to understand how “real-time” in real-time risk detection models were defined and operationalized in the reviewed papers. All papers in this review accounted for and discussed the timing of risk detection when presenting their approach. However, our review revealed that real-time or timely detection had different definitions, ranging from preventative risk detection to early risk detection. As shown in Table 2.2 and Figure 2.2, most of the papers (94%) presented early risk detection approaches to retrospectively detect the risks after it was posted online; yet, differed mainly in terms of the focus. Over half of the papers (51%), which form the majority of early approaches, focused on detecting the risk using *early stages* of information propagation or interactions (e.g., when only the initial retweets or comments became available). These approaches were trained to learn the risk using partial information of online content, such as the comments within the first 1-, 3-, and 5-day period (176). The rest of the papers considered the optimization of the least number of observations needed to make an accurate decision as part of models’ learning. A common challenge among these papers was how the model could achieve high accuracy with the lack of sufficient cues.

Another trend for the definition of “real-time” in the real-time risk detection papers revolved around the *detection time* to be as “early-as-possible” after the risk already occurred, without considering limited online content when training the models. Therefore, Hence, the latency of risk detection, the time gap between when a risk is detected and when it’s posted online, became a crucial metric. This measure was compared to the model’s accuracy over time to showcase the trade-off between early detection and accuracy. The improvement in model accuracy came at the expense of early detection, implying that the more data the models used to

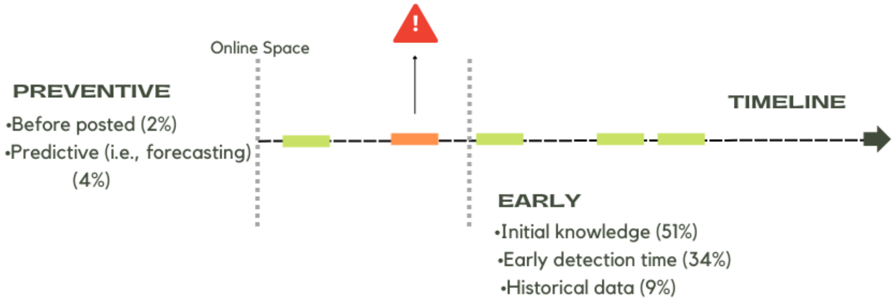


Figure 2.2: Timing in real-time risk detection approaches.

give accurate predictions, the more time the model would take to provide predictions. The third group of early detection methods expanded their input scope by incorporating *historical data* when detecting the risks early. As mentioned earlier, mental health indicators often depend on the presence of symptoms over an extended period. Consequently, we observed that the majority of papers utilized historical online content, such as changes in emotions (307), to the detection of mental health issues.

In this review, we found that sometimes, real-time risk detection was defined as a “preventative” approach that attempts to prohibit the risk from reaching online platforms (6%). The preventative approaches could be divided into two types: 1) prevent the risk offline prior to being posted online (2%) (223) or 2) predicting the possibility of risk occurrence in the future (4%) (202; 92). In both preventative approaches, the main goal was to prevent the risk from reaching online spaces, attempting to reduce its possible harm. For the first type of preventative approach, Masud et al. (223) presented a normalization real-time model for hate speech that intervenes when users type hateful keywords (i.e., an auto-complete fashion) and suggests normalized texts as an alternative before toxic words are posted online. This approach aimed to encourage individuals to post less toxic opinions online by providing proactive sensitization to them.

The second type of preventative approach (i.e., forecasting) formulated the risk detection problem for a given post and its initial history of comments, the model should forecast the risk, in the upcoming comments. Unlike early approaches that relied on delays to measure the models’ earliness, forecasting approaches relied on measuring the leading time for the model to accurately predict future risk incidents. Specifically, the effectiveness of these proactive approaches was measured by how accurate the model was in predicting the risk within N number of future comments. For instance, Dahiya et al. (92) evaluated the foreseeability of the hate intensity model by illustrating how far the model can predict hateful comments for $n = 1, 3, 5, 7$ and showed that the model performed well even for 7 future comments with Mean Absolute Percentage Error (MAPE) of ¡ 40. Forecasting or predicting social media risks can be useful to prevent risks from reaching online spaces and reduce damage. However, the applicability of such predictions was found to be more challenging than the early approaches as it is difficult to determine the exact occurrence and impact of risks beforehand, making it challenging to evaluate the effectiveness of forecasting models objectively.

2.4.2 Applying a Human-Centered Lens to Computational Trends for Real-Time Risk Detection

We organize our results by the dimensions of Razi et al.(290)'s framework to highlight trends and best practices in HCML. Then, we present our findings from the 53 articles analyzed in parallel structure to our codebook in Table 2.1.

2.4.2.1 Characteristics of the Datasets

The HCI and HCML communities care a great deal about leveraging ecologically valid datasets that are representative of the target populations they set out to study (319; 290). Given that data size and type are the foundation of algorithmic development, the HCML lens emphasizes making sure that the datasets match the real-world users' context (149). From a human-centered perspective, collecting ground truth annotations from humans, specifically from actual victims ensures that the training risk detection models reflect real-world experiences and accurately represent the risks users face online (179). Furthermore, leveraging human insights and theories when building real-time risk detection for social media is crucial for building human-centered models (34). Below, we summarize the trends we saw in the real-time risk detection literature compared to these best practices from the HCML literature.

Data Source, Privacy-Level, and Selection Criteria: All of the articles in this review relied on utilizing publicly available datasets. Over half of the papers (68%) used a scraped dataset from Twitter, followed by Weibo (26%). Thirteen papers (25%) used datasets from image and video-based social media platforms (i.e., Instagram (15%) and Vine (9%)). Reddit discussions (9%) were also utilized for real-time risk detection, while only two papers (4%) (370; 223) used a dataset from Facebook. This made the Twitter platform (Now 'X') the most dominant platform for datasets that were used to train and evaluate real-time risk detection models. Since most of the datasets were scraped from public social media posts (94%), data were scraped from unidentifiable users of the platform, without specifying users' target characteristics (e.g., profile characteristics). Only three (6%) papers (283; 190; 375) studied specific user groups instead of using a general query to collect social media data from any user, such as obtaining data from both depressed and non-depressed users. While they focused on collecting data from targeted groups, the selection criteria were still based on keywords and hashtags within the posts for identification. Table 1 in the appendix reports the datasets that the papers examined in this literature.

Dataset Size and Data Types: We found that 47% of the papers used a large social media dataset that consisted of more than 1 million instances with a maximum of 40 million tweets (197). In addition, another set of papers (25%) utilized medium-sized datasets with thousands of instances, which ranged from equal to or more than 10K instances (13%) to more than 100K instances (11%). We also found (28%) of the papers used equal to or more than a thousand instances, with a maximum of around 5K posts (92; 358; 213; 223; 205). All the papers reviewed relied on text and metadata (e.g., mainly the time of the post) datasets. Only 15% of the papers used datasets that included media data, such as images or videos derived from Instagram or Vine. However, the authors only used the textual features extracted from captions of media, media content, comments posted on the media, or meta information such as the number of likes and shares. For instance, Chelmis and Zois (72) only used the text of the extracted emotional cues from the Instagram pictures to train their early cyberbullying risk detection model while López-Vizcaíno (208) used the extracted textual features from the videos, including the nature of the video content and emotions.

Ground Truth: We found a noteworthy proportion of papers (72%) were based on existing labeled datasets that were ready for other researchers to use, which was illustrated in Table 2 in the appendix. Given the large scale of the collected data, most of these datasets were labeled using automatic approaches such as keyword or event searching (30%) or fact-checking websites (8%). A few of the ready-to-use datasets were labeled through researchers (4%), experts (6%), or crowdsourcing (8%). Of the research papers that we reviewed, 38% of the papers collected and labeled their dataset, most of which used human annotations (28%), including researchers (9%), psychology students or professionals (6%), crowdsourcing (6%), and experts (e.g., platform monitoring team) (6%). The rest of the papers relied on automated or rule-based approaches to label the dataset (9%). For example, (273) used the “Hateful Symbols or Hateful People” dataset (273) to label by checking if there was a hateful symbol and speech term in the tweet, it was labeled as harmful.

Class Distribution: Half of the papers (58%) leveraged unbalanced datasets while the rest (42%) used balanced datasets for their risk detection models (as illustrated in Table 3 in appendix). While the unbalanced distribution of classes reflects the realistic distribution of the risks in social media platforms as the online risk interactions are significantly less than safe interactions, this unbalanced distribution could yield severe consequences based on learner’s prediction bias toward the majority class (310). Only 11% of the papers presented an approach or discussed how to ensure the model fairness when using unbalanced datasets. These papers introduced improvements to the models’ equation to ensure the reliability of the model results such as using a modified misclassification costs ratios (247), adding Class-Balanced loss (91), and Focal Loss (198), which apply “a class-wise re-weighting scheme”, that were presented by Sawhney et al. (306). Liu et al. (206) took another direction by using the PU-learning approach (195), to learn from positive (P) and unlabeled (U) instances. This PU-learning framework identifies a sample of pseudo-negative instances from the unlabeled dataset and the classifier was then trained using these samples. The authors showed that the accuracy performance remained the same across the fully labeled balanced dataset and the unbalanced dataset, presenting promising results for future research to adopt.

2.4.2.2 Pre-Processing and Model Development

From a human-centered perspective, it is crucial for the computational models to be grounded in human knowledge and human theories (290; 34). This grounding ensures that these models can better understand and serve the social and psychological needs of individuals and society at large (66). In addition, human-centered models should provide local interpretability for individual predictions, making it easier for users to understand the model's reasoning for the detection decisions (200). In this section, we describe the computational trends of data pre-processing and model development for real-time risk detection.

Data Pre-Processing: Most of the articles (83%) used a "streaming-like" approach by segmenting the dataset into equal sizes of chunks of data that mostly were predefined either by a fixed time window such as (193; 213; 375) or a fixed number of posts such as (190; 383; 58) as listed in Table 4 in appendix. These chunks of data were fed to the training models sequentially to produce real-time risk detection decisions. Scholars in this literature have identified that segmenting the datasets is a limitation as these chunks are not representative of real-world peoples' interactions. Therefore, a few papers (17%) implemented sequential training by incrementally adding data (i.e., posts and comments) in chronological order as they were available online to mimic how the data was typically available in the real world, without any segmentation. In fact, Leiva and Freire (190) compared the "first n", which is the first n of messages that were concatenated to make predictions, and the "dynamic" setting, which was messages that were used as they became available to make the predictions when a confidence value reached a certain threshold. They found the dynamic approach with a 0.5 confidence value threshold outperforming the first n chunks of data, with a 0.05 early risk detection measure and 0.77 recall, illustrating that the dynamic approach could be considered the best practice in this field.

Features Selection: Upon reviewing real-time risk detection for social media literature, we found all papers (100%) relied on machine learning features, such as textual, network, user-based, or temporal features; among them, significant emphasis was given to leveraging textual (66%) features for detecting social media risks(Table 5). The textual features were found to be drawn mainly from the posts, comments, image captions, or video descriptions, which included text embedding, Linguistic Inquiry and Word Count (LIWC), term-frequency-inverse document frequency (TF-IDF), and Bag-of-Words (BoW). Meanwhile, 51% of the papers emphasized the importance of considering the social network contextual clues such as the social network (i.e., derived from users' relations) and conversational network (i.e., formed through users' retweets or comments for a given post), which found to improve the detection models comparing with comment streams (307). Domain-specific or theory-driven features were found in 32% of the reviewed articles. For instance, the real-time risk detection approaches to detecting mental illness mainly relied on prior psychological research such as suicide ideation on emotional reactivity (335), intensity (199), and instability (264) to identify a user's emotional spectrum over time. Theories such as the wisdom of the crowd and domain-specific measures such as degree of skepticism and susceptibility were leveraged in papers(246; 203; 193; 201) to harness the fake news and rumors detection algorithms. For interpersonal risks, we found a trend among the papers that mainly focused on using domain-specific features such as hate lexicons (e.g., Hatebase[3] and Luis von Ahn's Offensive/Profane Word List [4]) or counting the number of hateful instances (92; 202; 197). A few papers (11%) leveraged domain-specific user behavioral features that described user behaviors associated with certain risks such as their influence and role in rumors propagation (159) or extracting the digital "user footprint" of their abusive behaviors across multiple platforms (370).

Features Computation: Features in real-time risk detection models were learned in a sequential and incremental fashion. With the massive scalability of social media, computing these features was one of the challenges discussed in the real-time risk detection literature to provide scalable and timely risk detection solutions while maintaining sufficient accuracy. Most of the papers we reviewed (92%) used a straightforward approach: computing the features over time by doing a full rerun on the data as they become available (193). A few papers (8%) introduced approaches to reduce the feature computation timing by applying an attention mechanism to differentiate the importance of calculating the features from risk comments (197) or sorting features in increasing order based on their importance to make an early accurate decision based on the most important features (72). Dahiya et al. (92) took another direction by leveraging majorization-minimization algorithms (80), where the model computed the parameters only on the recently observed data, which led to faster processing. Another approach presented in the papers was that once the features were computed for the first set of input (e.g., comments) when new input was available, the features for the new input were calculated while reusing the previously calculated features, leveraging the incremental computation (142; 143). Unlike the models that perform a full run as each new data is available, this approach resulted in less re-computation overhead and would capture the naturalistic way of users' online interactions to provide timely risk detection.

Input Prioritization: In this review, we identified that most of the papers (96%) applied the real-time risk detection models by considering equal importance to classify all public conversations (posts and comments) without having a procedure to prioritize these data for detection. The significance of this procedure is mainly related to increasing the responsiveness of the detection approach to protecting people when needed. In addition, having less number of conversations or messages to classify or schedule the examined input for risk detection would lower the computational overhead for feature calculation, which in turn would produce

[3] www.hatebase.org

[4] www.cs.cmu.edu/~biglou/resources/bad-words.txt

faster risk decisions. Only two papers (4%) (282; 382) in our review identified and addressed this gap. Under the assumptions that “most media sessions are not bullying in nature, so not all media sessions need to be monitored equally”, Rafiq et al. (282) applied the resources on public media sessions (i.e., posts and associated comments) that were most likely to result in cyberbullying by presenting a Dynamic Priority Scheduler (DPS), which dynamically assigned high priority to sessions to be examined by the detection model and low priority to the ones that were postponed until new comments were available. This scheduler showed maximum responsiveness when it was compared with other traditional approaches. Zang et al. (382) applied a machine learning algorithm to calculate the False/True probabilities based on the initial features of the events (multiple tweets and replies about a certain event). The events with a high probability were assigned a True/False label and the false information events were moved to another step to be tracked for the final decision. By doing so, the final detection algorithm had a smaller set of possible false events to track them rather than inefficiently tracking all events.

Algorithms: Most of the real-time risk detection models addressed in the literature implemented either deep learning (60%) or statistical approaches (40%). Papers that relied on statistical theories mainly leveraged Markov models (11%), Bayesian model (9%), posterior probability (8%), and State Space models (4%) as illustrated in Table 6 in the appendix. Unlike the aforementioned traditional models that do not account for the sequentiality of data, deep learning models used commonly within real-time risk detection papers were expected and proved to be effective. Yet, off-the-shelf deep learning models were found to suffer when implemented in real-time risk detection, mainly because they can not account for the uneven time interval between responses or comments. Therefore, Sawhney et al. (306) proposed an approach that utilizes a monotonically decreasing function of elapsed time to transfer time into appropriate weights. Convolutional Neural Network (CNN)–based classifiers often generate complex and less interpretable representations of text. Therefore, works such as Liu et al. (206) improved CNN for fake news detection by adding a position-aware attention mechanism, which is an extension of the basic attention mechanism (237), which was used to learn how much attention should be given to the data points in the sequence.

For statistical models such as the Markov model, the risk detection problem was formulated as a sequential or time-series text data that was represented as a chain of posts/comments. The papers we reviewed presented their improvements to tailor these models for the types of risk that were tackled. For example, Li et al. (193) improved the standard Kalman Filter, which is a mathematical algorithm used for state estimation to achieve progressive detection through learning the temporal information of time-series data that arrive irregularly. A few papers (11%) discussed approaches where deep learning and statistical models were combined to foster the deep learning models’ capabilities to capture irregular conversation evolving nature. For instance, Dahiya et al. (92) utilized state-space models that were combined with deep-learning models, known as deep-state models, where there was a sequence of unknown states that were considered as learned features to represent the data, and then, at each time step, the model provided a probabilistic estimate of the future hidden states conditioned to all available data up to time.

2.4.2.3 Real-time Risk Detection Evaluation

HCML framework highlights that when evaluating computational risk detection models, it is imperative to look into the models with human-centered perspectives to understand whether the models could make accountable and fair decisions (290; 227). This assessment could be achieved by incorporating qualitative explanations that go beyond the known quantitative performance metrics. In addition, leveraging the human-in-the-loop approach is one of the important standards that should be incorporated in building human-centered algorithmic results (93). In the following section, we will provide a detailed explanation of the quantitative and qualitative assessments of the models' performance, as derived from the literature we have examined.

Detection Performance: Timeliness and accuracy are associated metrics in real-time risk detection literature. All papers in this review leveraged commonly known accuracy metrics, including F1-score (70%), Accuracy (53%), recall (58%), precision (51%), Area Under Curve (11%), and Root Mean Square Error (RMSE) (4%). As we explained previously, most of the papers reviewed in this paper focused primarily on training the models using predefined fixed chunks of data. As a result, the models' evaluation was also done using these chunks of data. The majority of articles evaluated the timeliness of the models (i.e., accuracy performance over time) based on chunks of a fixed number of posts (53%) and fixed time windows (21%). These chunks were fed to the models in chronological order to measure the models' accuracy performance across fixed chunks of posts or time windows. A few papers (21%) used the detection time to evaluate how timely is the model. These papers mostly built models that learned when to stop using a widely known problem in statistics called the Markov Optimal Stopping problem (318). Meanwhile, two papers (384; 371) took another direction by leveraging reinforcement learning to identify the optimal number of observations needed to make the detection decision, which is the most promising approach that could provide assessments about the models' performance when it is deployed in real-world applications.

Explainability: Our review revealed that (57%) of the papers presented explanations for the model performance beyond the timeliness and known accuracy measures (Table 7). These papers presented qualitative explanations of real-time risk detection, including qualitative analysis (32%), error analysis (13%), case studies (9%), and human evaluations (2%). Qualitative or error analysis was discussed in papers to further explain their models' performance such as (202; 206; 250). Due to the goal of implementing the models in social media, the papers have mainly focused on minimizing and discussing the false negatives (202; 206). For instance, Liu and Guberman et al. (202) found that the false negatives of their hostility forecasting mainly occurred when there was no indication of escalation with many consecutive similar innocuous messages sent. Case studies were leveraged to illustrate and explain how the model performed over time using a case. None of the reviewed papers integrated user studies or human evaluations, with the exception of Masud et al. (223), who surveyed 25 participants to assess their model that altered hateful texts and found that their model outperformed other hate normalization models in terms of generating reduced hateful comments and more fluent sentences. This human evaluation demonstrated that the effective performance of their hate normalization model extended beyond the dataset used during training.

2.4.2.4 Applications and Interventions of Real-Time Social Medial Risk Detection

HCML places a strong emphasis on building systems-based artifacts to foster machine learning transparency that allow humans to explore machine learning used features and decision results to build trust, making these models less of a "black box" and enhancing their usability and societal impact (333). The real-world artifacts should be designed to empower users to interact with, question, and comprehend the algorithms' inner workings to elevate stakeholders' oversight such as victims, clinical practitioners, and social media platform owners or moderation teams (290). The HCML community also promotes the development of risk intervention to ensure that machine learning models are designed to minimize harm and adverse consequences for individuals and society (171). Below, we identify the artifacts and risk interventions that were presented in the literature.

Applications: The majority of the papers (92%) focused on presenting the algorithmic approaches that enhanced real-time risk detection, yet presenting system artifacts or APIs that could be integrated with social media platforms was rarely done in the literature. Two papers (4%) presented an online-offline detection approach where the model is fully trained offline and the trained model was deployed in a social media platform or server hosting services (370; 92). Another two papers (223; 385) developed an interface to demonstrate the performance of their models in the real world. For instance, Zou et al. (385) developed a web interface in which users can search for an event, then an alert would appear if the event was likely to be a rumor along with three visualizations that illustrated 1) the event's timeline to show the event evolution along the time deployed, 2) the propagation structure on social media, and 3) user information graph. Only one paper by Rafiq et al. (282) conducted a simulation for their model in Amazon AWS virtual machine instances with 1GB memory to evaluate the scalability of their model by replicating the 100,000 media sessions' traffic up to the scale of 39 million sessions.

Interventions: The majority of the papers we reviewed (89%) focused on the detection algorithms and their performances. However, a few papers (11%) presented an intervention strategy such as alerts, alternating the risk language in posts, or immunizing certain users from receiving risk content. Three papers (6%)

presented alerts that were raised when cyberbullying instances were detected (374; 282; 72). Intuitively, these alerts should be raised after a classifier produced a decision; however, these papers discussed waiting until certain positive decisions have been made to avoid false positives, which was identified as a trade-off between responsiveness and precision. For example, Yao et al. (374) introduced an approach that reviewed Instagram comments as they became available over time and raised an alert only when the total number of comment–level detection decisions topped a certain threshold. Two papers (4%) (273; 368) leveraged the network immunization approach with the goal of minimizing the spread of risk information such as hate speech or rumors. This approach is mainly derived from network science and graph theory to identify these nodes or users effectively. For instance, Petrescu et al. (273) utilized preventive immunization, which worked on the network without knowing the source of risk content and was applied after detecting hateful content, by lowering the rank of that particular post in the feed. As such, we have identified prior efforts to advance the state-of-the-art of real-time risk detection approaches. Next, we will briefly describe the human-centered gaps and recommendations to direct future research to the best practices in this field.

2.5 Discussion

This section describes the identified gaps along with the recommendations to address them and advance the real-time risk detection approaches computationally and from a human-centered perceptive.

2.5.1 Identifying the Gaps in Real-Time Risk Detection Research from the Human-Centered Perspectives

First, our analysis provided an opportunity to extend Razi et al.'s framework for systematic reviews of computational risk detection literature by adding unique dimensions for human-centered perspectives of real-time risk detection. The new dimensions and codes that emerged included characteristics of the **dataset** (i.e., selection criteria, dataset size, class distribution), **pre-processing and model development** (i.e., data processing, feature computation, input prioritizing), and **evaluation** (i.e., timeliness). This methodological contribution is valuable for future systematic and human-centered reviews of computational risk detection literature that involve real-time approaches. In this section, we describe the gaps in the social media real-time risk detection literature (illustrated in Table 2.3) and how to address them from a human-centered perspective moving forward.

2.5.1.1 Datasets Gap: The Absence of Ecologically Valid Datasets

We raise several questions regarding the ecological validity of datasets for real-time online risk detection. The current approach heavily relies on publicly available text datasets scraped from platforms, excluding input from humans, victims, or survivors of these risks at any stage. Depending solely on such data could hinder the effectiveness of real-time online risk detection. Moreover, collecting the data and ground truth annotations from humans, specifically from actual victims ensures that the training risk detection models reflect real-world experiences and accurately represent the users and risks they face online (23). Further, in section 4.1.1, we described the *time-evolving nature of risks*; relying solely on static public datasets might overlook the nuanced dynamics inherent in how these phenomena unfold over time. We note the need for capturing temporal patterns, such as escalation during cyberbullying, the gradual unfolding of mental health symptoms, or the trust-building stages that have been well-documented for sexual grooming (63), which require longitudinal data for timely and accurate identification. Additionally, in fake news and rumor detection, acknowledging the progress of content spread could be crucial for collecting datasets that represent the dynamic nature of these risks and associated human behaviors. Therefore, we recommend considering data collection methods and advanced systems designed to gather real-time and continuous data streams, or at least robustly simulate interactions that occur over time. This approach should be tailored for specific populations (e.g., risk victims or survivors), the actual contexts of risks, as well as the dynamic aspects of risk escalation and human communication.

2.5.1.2 Models Gap: The Need for Grounding Models with Human Behaviors

Our analysis revealed that the existing models were grounded on primarily computational efficiency considerations, without considering human understanding or theories. Most papers used a streaming-like data processing approach with data chunks lacking conversation context, which could lead to the model misinterpreting or missing potential risks. Only 32% of the papers developed features based on human theories and domain-specific knowledge to capture nuanced context. Therefore, we highlight the significance of acknowledging the dialectical nature of human communication and the dynamic changes in behavior within risk contexts when designing features to enhance the effectiveness of real-time risk detection algorithms. This acknowledgment emphasizes the need for designing online features that capture sequential conversational data rather than traditional (i.e., all conversation at once) or chunk-based features (357). In addition, well-established methodological approaches like discourse analysis (54), which provide a foundation for in-depth exploration of the structural aspects of human communication, could be useful to craft these features by identifying the time-evolving nature of human communication such as shifts in tone, frequency of aggressive language, shifts in mood or self-disclosure, or changes in narrative. Incorporating such approaches into the design of algorithms enables a more nuanced interpretation of online interactions that aligns closely with human understanding. In this review, we also found heavy reliance on the high capability of deep learning models; yet, these models were not inherently human-centered since these models often operated as “black boxes." This can hinder stakeholders, including the victim, from understanding the models' output (330). Therefore, there is still a need to adopt human theories widely and human-centered real-time risk detection that effectively identifies social media risks.

2.5.1.3 Evaluation Gap: Involving Humans in the Evaluation Process is Needed

We found most of the papers relied on purely computational metrics (e.g., accuracy and timeliness) without incorporating user studies or human insights into the evaluation of developed risk detection models and identifying the effectiveness of the models' timeliness in protecting people. We even conducted additional searches for subsequent user studies related to the reviewed papers, but we only found one publication by Chang et al. (68) in which Liu's et al (202) hostility forecasting model was embedded into a tool assessed by end users. Their data collection included a survey on participants' experiences with incivility, responses to tool warnings, and overall impressions, alongside real-time recording of drafting behavior via usage log data. They found that the proactive incivility warnings enhanced participants' awareness of their interactions by reflecting more on conversation tension, spending more time drafting comments, and revising replies to mitigate any tension. Similar to this research, future real-time risk detection models could consider incorporating human evaluations to ensure that these models align with human values, ethics, the complexities of online communication, and aligned with evolving risk dynamics, ultimately leading to more effective, trustworthy, and responsible models. However, we recognize the complexities involved in carrying out these evaluation studies concerning ethical considerations, especially those related to algorithmic bias (68). These issues present difficulties in mitigating potential negative impacts like the reinforcement of stereotypes or the marginalization of vulnerable groups, as noted by Xu et al. (370). Despite these challenges, exemplars like the user study by Chang et al. (68) have established a pathway for future researchers to navigate and potentially tackle these ethical and technical concerns in conducting user studies to evaluate risk detection algorithms. Consequently, we emphasize the need for collaborative initiatives to engage in ethical discussions, aiming to identify the best practices for conducting such important user studies.

2.5.1.4 Applications and Interventions Gap: Need for More Real-Time Interventions and Real-World Applications

Most papers focused more on presenting effective detection algorithms without presenting system-based artifacts and interventions using real-time risk detection algorithms. The existence of such systems is a necessary prerequisite for research on real-world algorithm deployment, system design, and user experience resulting from the use of such systems. These studies are important to improve our understanding of how users engage with and react to applications designed for risk detection. In fact, deploying risk detection models in real-world applications has become more of an industrial problem than an inherent expectation in research presenting these detection algorithms (179). Moreover, the availability of open-source risk detection systems or algorithms is limited, often confined to proprietary platforms or academic papers (61). Therefore, HCI scholars could bridge this gap by redirecting the fields' attention and resources toward developing interfaces and interventions. Additionally, fostering interdisciplinary collaboration between experts in ML and HCI fields could lead to the development of such systems and algorithms with the goal of aligning them with user expectations. Building artifacts should be designed to empower users, including victims, clinical practitioners, and social media platform owners, to interact with and understand how these algorithms work. When stakeholders can explore predictions, understand decision factors, and question the algorithm's outputs, they can intervene if needed to align with human values and privacy considerations, improving the algorithm together. For instance, employing personalized interventions could play an important role in offering targeted support based on individuals' preferences, needs, and behaviors, while empowering users with a sense of control and autonomy (73). As a result, these artifacts would foster real-time risk detection models' transparency, making the algorithms less of a "black box" and enhancing their usability and societal impact.

2.5.2 Establishing a Research Agenda for Real-Time Risk Detection on Social Media

We make several recommendations for advancing real-time risk detection approaches based on our review. Figure 2.3 illustrates our conceptualized and comprehensive framework to direct future research with recommendations for the best technical and human-centered practices for real-time risk detection algorithms.

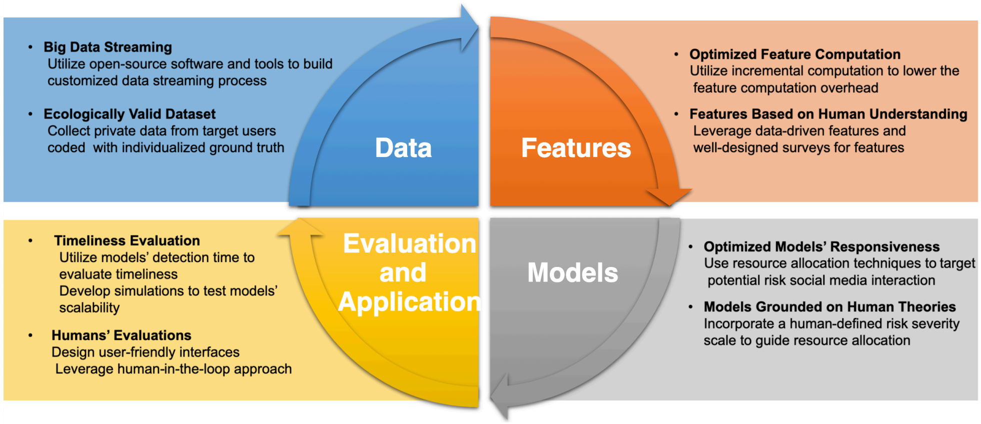

• Big Data Streaming
Utilize open-source software and tools to build customized data streaming process
• Ecologically Valid Dataset
Collect private data from target users coded with individualized ground truth
Data
Features
• Optimized Feature Computation
Utilize incremental computation to lower the feature computation overhead
• Features Based on Human Understanding
Leverage data-driven features and well-designed surveys for features
• Timeliness Evaluation
Utilize models' detection time to evaluate timeliness
Develop simulations to test models' scalability
• Humans' Evaluations
Design user-friendly interfaces
Leverage human-in-the-loop approach
Evaluation and Application
Models
• Optimized Models' Responsiveness
Use resource allocation techniques to target potential risk social media interaction
• Models Grounded on Human Theories
Incorporate a human-defined risk severity scale to guide resource allocation

Figure 2.3: Recommended Computational and Human-Centered Framework For Real-Time Risk Detection Approaches for Social Media

2.5.2.1 Towards Leveraging Streaming Mechanisms for Ecologically Valid Datasets

We propose that real-time risk detection training and testing could eventually be accomplished using private and multimedia streaming data or online processing of continuous data, instead of relying on predefined chunks of data. Social media environments are characterized by rapidly changing data patterns, influenced by user behaviors, trends, and external events; therefore, continuous data streaming systems could capture this dynamism (168). Developers of real-time risk detection are encouraged to construct personalized data stream processing systems utilizing open-source software and tools such as Kafka, rather than using commercial or proprietary systems that may not adhere to the users' privacy (162; 157). Future research in real-time risk detection could also leverage informative reviews on data streaming systems such as (162; 168) that provide insightful information about the usability, features, and real-world use case scenarios for different data streaming systems.

From a human-centered perspective, to ensure that training models reflect the real-world experiences of users, these data streams could be obtained directly from victims or survivors of online risks in human-subject studies with their consent. We acknowledge that data collection from such vulnerable populations is uniquely challenging as it requires researchers to ask them to share and label their intimate online conversations while ensuring that participating in research does not harm them (249; 287). We suggest that researchers set guidelines beforehand and make sure to follow established recommendations to enhance the ethical implementation of research involving survivors and victims of traumatic online risks, for example by putting in place formal consultation procedures for participants (88). In addition, scholars in the HCI field have initiated efforts such as MOSafely[5] (Modus Operandi Safely) with the objective of establishing a multidisciplinary collaborative community that concentrates on safeguarding young individuals in online environments. This innovative approach may serve as a potential avenue for resource-sharing, encompassing datasets and algorithms, to effectively address the online security concerns pertaining to at-risk youth. It is advisable for future research endeavors to actively participate in such collaborative initiatives with a strong ethical foundation, prioritizing the protection of the rights and well-being of people.

2.5.2.2 Toward an Optimized Real-Time Models' Efficiency Grounded in Human Understanding

In this review, most papers leveraged an aggregate view of features over time or time windows. By aggregating the features, the models consider them as independently distributed, meaning that the features calculated for one set of data are unrelated to newly available data, which fails to capture valuable information from adjacent time periods (i.e., evolutionary data) (151). Due to these reasons, we suggest that an optimal solution for calculating the features could account for the computation overhead. We suggest that the best practice

[5] https://www.mosafely.org/

identified to address this issue in this review revolved around calculating the features for a set of data, and then when new data became available, the features of this new data were only calculated while the previously calculated features were reused, which has mainly been implemented through incremental computation (282). This approach proved to reduce the computation complexity of features calculation and provide faster classification than classical approaches. Therefore, setting a benchmark for the performance of the incremental computation in features engineering and balancing between the models' computation efficiency and accuracy is a task for future research in real-time to further investigate.

Unlike ML scholars who mainly leveraged the data-driven features for real-time risk detection, socio-psychological researchers often employ survey-based or interview methods to capture contextual information directly from human subjects. However, in both cases, the data may not provide a holistic understanding of human experiences or behaviors that could be helpful for real-time risk detection. Therefore, we call ML and socio-psychological researchers to collaborate on designing a complementary approach to utilize data-driven features that provide objective insights about users' experiences and subjective data collected through well-designed surveys to provide a holistic understanding of human experiences, behaviors, or perspectives of risk. Additionally, prior research has shown that context-based features improve the detection accuracy performance (326), and this is particularly relevant to real-time risk detection on social media where users' interactions could transform within minutes. Therefore, the dynamic updates of user behaviors ensure that the prediction model remains reflective of the shifting patterns of these behaviors (328). Yet, capturing this time-evolving context in terms of scale becomes challenging and has been identified as a crucial avenue for future development (12). Therefore, future research in real-time risk detection approaches could further investigate the applicability of incorporating such features, monitoring how this will affect the models' scalability.

2.5.2.3 Towards Advancing Real-Time Risk Detection Responsiveness and Interpretability

We identified two papers that adopted a procedure to enforce the risk detection models targeting potentially risky social media interaction (i.e., priority schedulers and machine learning-based ranking) (282; 382). These papers pave the way for future real-time risk detection algorithms to be more responsive by allocating resources to focus on conversations that are more likely to require immediate intervention. However, we also suggest exploring and using other resource allocation techniques such as adaptive allocation to continuously monitor the workload of the risk detection system and allocate resources based on the volume of conversations and the urgency of risk detection, or predictive allocation to anticipate periods of high-risk activity based on historical patterns to allocate resources during these periods. Reinforcement scheduling (33) leverages reinforcement learning that could be used to learn when potentially risky conversations or at-risk populations would need the detection algorithms. Therefore, future research is encouraged to adopt these proactive allocations of resources to effectively target conversations or threads that potentially contain risky content and ensure faster responses to emerging risks, creating a safer online environment for social media users.

An ultimate approach to ground these recommended resource allocation techniques with human understanding could be by incorporating a human-defined risk severity scale; therefore, more efforts to understand the severity of online risk from the perspectives of users (e.g., (362)) are needed. One approach could be identifying profiles of at-risk individuals that might need more attention from the risk detection algorithms, using unsupervised clustering techniques (324) or by leveraging survey-based data to feed well-established statistical techniques (e.g., Mixture Factor Analysis (244)). Future research on improving the real-time risk detection algorithms is warranted to leverage such human-centered practices when optimizing the models' responsiveness. To achieve this goal, we urge HCI scholars to collaborate with ML and Data Scientists to guide the resource allocation process based on human understanding. In addition, Our findings inform that combining deep learning models with state models can help incorporate domain-specific constraints and handle uncertainty more effectively. Therefore, we recommend future researchers to investigate combining deep learning and statistical models in an ensemble approach, and how it could impact the models' interpretability. Since ensemble methods aggregate predictions from multiple models, they can capture complex patterns while benefiting from the transparent insights of statistical models.

2.5.2.4 Towards Designing Applications to Incorporate Human Evaluation and Personalized Interventions

In our review, we found that reinforcement learning (RL) such as Q-learning or deep reinforcement learning has the most promising potential to provide information when the detection decision was made instead of relying on pre-defined chunks of data (196; 323). These techniques advance the detection models to know what level of cues is enough for the model to review the input and provide the detection decision. Besides the detection time, another performance measure should be considered: how well the detection models perform as data volumes increase using a nearly realistic environment. Incorporating scalability simulations into the evaluation process is crucial for ensuring that real-time risk detection models can effectively handle the dynamic nature of social media data streams. During these simulations, it is important to identify potential bottlenecks, limitations, and performance degradation in detection latency and compute times that are essential to enhance the overall responsiveness. Therefore, we recommend that future research on real-time risk detection provide performance metrics that are more useful when deploying the models in real-world settings.

To fill the gap between technical solutions and human expectations, a growing body of work has highlighted the importance of human insights into algorithmic performances to facilitate HCML by informing developers entrusted with designing ethical machine learning algorithms and decision-makers tasked with implementing such systems in social contexts (175). Given explainability and fairness perceptions are highly context-dependent and can vary substantially across domains, tasks, and algorithmic designs (327), human involvement in evaluation processes is essential. One way to reflect human perceptions in the evaluation of machine learning systems is through interactive machine learning system design (19) in which human end users are iteratively involved in the model development process. Participatory design strategies allow the users to learn about how the machine-learning model works by instantly testing various inputs and examining the impact of the models (356; 110; 364). More importantly, these user-led cycles of trial-and-error discovery processes can help developers steer the model to improve model outcomes in ways that satisfy those who are affected by the models the most. Therefore, we call for more collaborative approaches among multi-stakeholders including developers, designers, and users to work together through co-design sessions (6), or even more long-term efforts such as the advisory board of users (241). This way, we can make sure that real-time risk detection models are working in ways that meet users' expectations and benefit those who are affected by online risks.

Designing user-centric or personalized interventions could involve multiple steps to ensure the effectiveness of these interventions. First, researchers are recommended to gather data to identify the target users and create profiles of users, which include scraping social network data and self-reported to understand individuals' needs, behaviors, and preferences. Scholars have also called for going beyond individual characteristics to explore the effectiveness of contextual characteristics such as culture (294). Additionally, these interventions should be adaptable to the evolving nature of users' behaviors and needs by continuously monitoring user interactions and feedback to ensure that the support provided remains relevant and engaging. Nudges or gamification could be integrated with these interventions to improve the overall user experience (4; 49). The design of personalized interventions should possess visual attractiveness, simplicity, and personal relevance in order to resonate with any particular population (253). Future researchers are encouraged to collaborate with interdisciplinary teams such as psychologists, user experience designers, and data scientists to ensure that intervention designs consider psychological, technical, and ethical dimensions.

2.5.3 Reconsidering 'Timing' in Real-Time Risk Detection

We found that 94% of the papers operationalized real-time risk detection tasks as an early detection approach that worked on the data retrospectively to detect the risk as early as possible. Social media-rich data have been proved in prior studies to be successfully used to predict the future (i.e., forecasting) across different domains and contexts such as marketing, finance, and sociopolitical events (299). However, despite our adoption of a comprehensive view of online risks in contrast to prior reviews (290; 179), we observed a similar trend in terms of a conspicuous dearth of preventive methodologies with respect to risk prediction and mitigation. A possible explanation of this rare implementation could fundamentally arise from challenges such as data noise, biases inherent in social media data, limited generalizability, and the inherent difficulty in integrating domain-specific knowledge and theoretical frameworks (274). In addition to these identified challenges, in our review, we observed that the rapid dissemination of information on social media frequently resulted in temporal intervals that are insufficiently extensive for models to anticipate and proactively address emerging risks before they materialize or escalate. With that being said, we presented three research papers (202; 92; 223) as exemplars that future scholars may consider when seeking to apply and explore the efficacy of their novel preventive methodologies across diverse datasets and various risks.

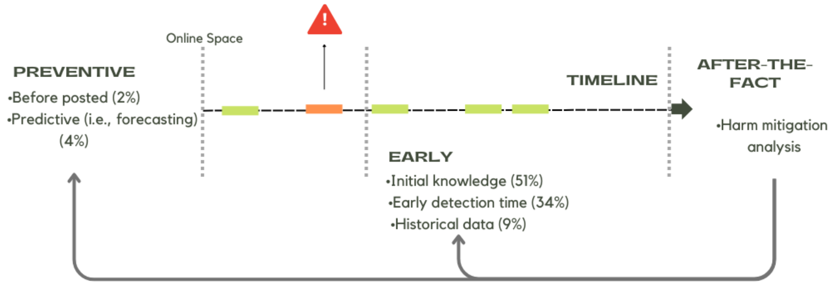


Figure 2.4: Comprehensive real-time risk detection approaches.

Another interesting finding in our review was the trade-offs identified within early risk approaches between accuracy and latency in ways that the more data the models use to give accurate predictions, the more time the models take to provide predictions. Trade-offs in ML-based computational systems have been well-documented in the literature, especially between fairness and accuracy, which is a value-sensitive and open question for further discourse (269; 268). We highlight that striking a balance between accuracy and timely detection is indeed an important and challenging aspect of real-time risk detection, especially given that real-time risk detection models are designed to provide "just-in-time" intervention to support those who are (potentially) at risk. Hence, careful consideration of how to balance the two is essential for future work toward designing value-sensitive and effective computational systems to support individuals and society. One way to accomplish this balance might be by defining acceptable trade-off thresholds between accuracy and latency. For example, accepting a certain drop in accuracy if it significantly reduces latency. Thus, the optimal balance between accuracy and latency will vary based on the specific use cases and requirements of the detection task.

In fact, preventive and early approaches aim to safeguard people on social media platforms from potential harm, yet they differ in terms of timing and focus. On one hand, early detection is rooted in real-world data, which could lead to more accurate risk assessments than preventive approaches. Yet, the time required for detection, analysis, and response may result in a delay between risk emergence and effective intervention, reducing its efficacy within a rapidly evolving environment such as social media platforms. On the other hand, the preventive approach utilizes the predictive indicators to take action "before" the risk incident occurs or the victim suffers from the risk. However, applicability concerns have been posed about this approach as explained previously that may lead to unnecessary content removal or user restrictions if these models were not trained very well (70). As such, each approach (i.e., preventive and early risk detection) has pros and cons that are warranted to be balanced in future research. We suggest that ultimately, a combination of both strategies along with late risk mitigation, tailored to the specific context and nature of risks, can be the most effective way forward in building a safe online landscape, as illustrated in Figure 2.4. This means that preventive, early, and late risk mitigation strategies could be developed hand in hand to provide a comprehensive risk detection approach that detects the risk as early as possible in case the predictive indicators fail to forecast and mitigate risks beforehand. Late risk mitigation could serve as an analysis stage of the risks that were missed by the preventive and early approaches or the risks' long-term impact (e.g., cyberbullying and following mental health indicators). Adopting this approach forms a full cycle of real-time risk detection algorithms to effectively ensure individuals' safety.

2.5.4 Limitations and Future Research

There are several limitations of our review that are worth mentioning. First, while our review was comprehensive, it is possible that we did not include all published work that met our inclusion criteria. Additionally, we limited our inclusion criteria to papers that developed a novel computational real-time risk detection model for social media-related risks. Considering the computational complexities involved in developing and assessing these algorithms, it is probable that the human-centered evaluations of these systems were left for subsequent work, although we did not find many in this vein. Consequently, we strongly encourage more research focused on HCI aspects of real-time risk detection on social media, including intervention-based approaches, interface design, user experiments, and real-world system deployment. Further, there may have been some papers that met our inclusion criteria that were held out-of-scope because it was difficult for us to evaluate relevancy due to inconsistent reporting standards. Therefore, we urge the HCI and ML research communities to converge on local norms for reporting important metrics uniformly across fields to increase the communities' ability to synthesize the results in a way that moves the fields cohesively forward. Furthermore, this review primarily concentrates on peer-reviewed research, yet it is worth noting that numerous social media companies are independently developing proprietary algorithms for real-time risk detection (116). To advance the field more effectively, fostering collaboration between academic and industry researchers could prove to be highly advantageous. Finally, all human-centered research is nuanced, complicated, and context-dependent.

As such, insights regarding specific risk types may not be directly applicable to other risks, especially when comparing interpersonal risks, such as cyberbullying or mental health to community-level risks, such as fake news. Therefore, future researchers should use their discretion, as well as domain experts' opinions, as to what recommendations make sense in the context of their work.

2.6 Conclusion

In an increasingly digitalized society, individuals face growing complexity due to the diverse range of social media risks, impacting both individuals and society as a whole. Detecting these risks accurately and timely has become a pressing necessity to facilitate effective interventions for various stakeholders, including governments, online platforms, societies, and academic communities. While previous studies have made great progress in advancing real-time risk detection approaches for social media, our review revealed a lack of integration with human understanding and behaviors in these approaches. Therefore, we strongly recommend that future research prioritize placing humans at the center of designing, developing, and testing real-time risk detection systems to ensure their effectiveness in real-world settings. As our review highlights, as HCI researchers, it is imperative for us to join forces with ML developers and researchers to bridge the gap between theoretical socio-psychological knowledge and the hands-on implementation of computational solutions for real-time risk.

Table 2.1: Codebook for RQ2 RQ3 ($N = 53$ articles) based on the Razi et al. framework (290) for performing a human-centered review of computational research. **Note:** * and bolded text in the table represents new dimensions and/or emerging codes we added to Razi et al.'s framework to extend it to better account for research that focuses on 'real-time' computational risk detection.

Razi's et al. (290) Dimensions	**Codes**	**Sub-Codes**
Characteristics of the Datasets: *What were the sources for data collection? What was the privacy level of the dataset? Was the data collected from targeted users? How large were the datasets? What were the data types? How was the data annotated for training datasets? What was the distribution of classes?*	Data Source	Twitter (68%), Weibo (26%), Instagram (15%), Vine (9%), Reddit (9%), Meta (4%).
	Privacy Level	Public (100%), Private (0%)
	Selection Criteria*	Unidentifiable users (94%), Targeted users (6%)
	Dataset Size*	Large (47%), Medium (25%), Small (28%)
	Data Type	Text (100%), Meta (100%), Images (11%), Videos(4%)
	Ground Truth	Existing (74%), Third-party annotators (26%), Automatic (17%)
	Class Distribution*	Balanced (42%), Unbalanced (58%)
Pre-Processing and Model Development: *How was the data processed for simulating real-time? What were the features and how were they calculated for the model? How the data were prioritized to review and detect risk? What machine learning model (s) were used?*	**Data Processing***	Fixed chunks of data (83%), Dynamic input (17%)
	Feature Selection	Domain specific/ Theory Driven (32%), General ML features (100%)
	Feature Computation*	Straightforward (92%), Optimized (8%)
	Input Prioritizing*	Equal prioritization (96%), Prioritizing technique (4%)
	Algorithms	Deep learning (60%), Statistical (40%)
Evaluation: *What accuracy and timeliness metrics were used? What explainability analysis was incorporated to explain the models' performance?*	Accuracy	F1-score (70%), Accuracy (53%), Recall (58%), Precision (51%), AUC (11%), RMSE (4%)
	Timeliness*	Fixed chunks of input (53%), Fixed time window (21%), Time (21%)
	Explainability	Qualitative analysis (32%), Error analysis (13%), Case studies (13%), Models' fairness (2%), and Human evaluations (2%)
Application and Interventions: *What were the final artifacts? What interventions were provided for risk mitigation?*	Applications	Algorithm only (92%), Interfaces (4%), Deployment (6%)
	Interventions	Alerts (6%), Immunization (4%), Language alteration (2%)

Table 2.2: Real-Time Approches

Real-Time	**Types**	**References**
Early (94%)	Initial Knowledge (51%)	(58; 72; 75; 98; 159; 176; 194; 193; 197; 205; 206; 208; 215; 217; 232; 246; 247; 250; 297; 304; 305; 322; 369; 370; 374; 383; 384)
	Early Detection Time (34%)	(78; 77; 114; 129; 201; 203; 213; 273; 282; 283; 350; 358; 368; 371; 375; 378; 382; 385)
	Historical Data (9%)	(190; 306; 307; 367; 387)
Preventive (6%)	Before Posted (2%)	(92)
	Predictive (i.e., forecasting) (4%)	(202; 223)

Table 2.3: State-of-the-art in real-time risk detection computational approaches and the identified human-centered gaps.

	State-of-the-art Computational Approaches	**Gaps from the Human-Centered Lens**
Dataset	Utilized large-scale, public datasets with established ground truth.	The absence of ecologically valid datasets that are representative of targeted population and the contexts of their online risk experiences.
Models	Trained models using streaming-like data, textual and social network features, and improved deep learning.	Lack of grounding pre-processing, features, and models with human behaviors.
Evaluation	Evaluated the models' using chunks of data for timeliness and qualitative and error analysis to interpret the models' performances.	Lack of human evaluations of the models' performance.
Applications	Presented novel algorithmic approaches for real-time risk detection.	Lack of artifacts to deploy the models in real-world settings and personalized interventions to intervene after detection.

CHAPTER 3

STUDY 1: From ‘Friends with Benefits’ to ‘Sextortion:’ A Nuanced Investigation of Adolescents’ Online Sexual Risk Experiences

Citation: Ashwaq Alsoubai, Jihye Song, Afsaneh Razi, Nurun Naher, Munmun De Choudhury, and Pamela J. Wisniewski. 2022. From ’Friends with Benefits’ to ’Sextortion:’ A Nuanced Investigation of Adolescents’ Online Sexual Risk Experiences. Proc. ACM Hum.-Comput. Interact. 6, CSCW2, Article 411 (nov 2022), 32 pages. https://doi.org/10.1145/3555136

Based on our literature review findings, I aimed to identify the myriad of youth online risk experiences. We started with sexual risks as it was found to be one of the most prevalent risks that youth encounter online. Sexual exploration is a natural part of adolescent development; yet, unmediated internet access has enabled teens to engage in a wider variety of potentially riskier sexual interactions than previous generations, from normatively appropriate sexual interactions to sexually abusive situations. Therefore, we analyzed posts ($N = 45,955$) made by adolescents (ages 13–17) on an online peer support platform to deeply examine their online sexual risk experiences. By applying a mixed methods approach, we 1) accurately (average of $AUC = 0.90$) identified posts that contained teen disclosures about online sexual risk experiences and classified the posts based on level of consent (i.e., consensual, non-consensual, sexual abuse) and relationship type (i.e., stranger, dating/friend, family) between the teen and the person in which they shared the sexual experience, 2) detected statistically significant differences in the proportions of posts based on these dimensions, and 3) further unpacked the nuance in how these online sexual risk experiences were typically characterized in the posts. Teens were significantly more likely to engage in consensual sexting with friends/dating partners; unwanted solicitations were more likely from strangers and sexual abuse was more likely when a family member was involved.

3.1 Introduction

Easy access to these online platforms has enabled adolescents to express and explore their sexuality in new ways (340; 288). “Sexting,” or engaging in sexual conversations, sharing flirtatious comments, or sending sexual self-images online has been a topic of research inquiry for over a decade (329; 191). More recent literature (105; 243; 97) has broadened the definition of sexting to also include sending, receiving, and forwarding any kind of sexual messages (e.g., text, images, videos) across various technology-mediated platforms (e.g., text messages, email, social media). Although adolescent sexual exploration is considered developmentally normal (277), public discussions of adolescents’ online sexual behavior have focused primarily on its possible risks, harm, and detrimental consequences. Research on adolescent sexting has focused on examining its association with adverse consequences, such as substance use, risky sexual behaviors, coercion, anxiety, depression, and suicide (29; 71; 376; 52). However, other researchers have found no association between adolescents' sexting and negative consequences (160; 240). Therefore, a broader perspective of adolescents' online sexual experiences should be explored by going beyond treating these experiences as either *risky* or *safe* and by contextualizing these online sexual experiences in a nuanced way (245). Prior research has shown that youths' online sexual risk experiences vary significantly based on whether the interaction was consensual or non-consensual (102; 353) and on whether the relationship was between intimate partners or strangers (59; 153). We conducted an in-depth examination of teens' (ages 13–17) disclosures about their online sexual risk experiences to address the following research questions:

- **RQ1:** *Can adolescent online disclosures about their online sexual risk experiences be accurately identified from their posts? If so, can these posts be further classified by a) level of consent and b) relationship type?*

- **RQ2:** *Are there distinguishable differences and/or patterns in adolescent sexual risk experiences based on level of consent and relationship type?*

- **RQ3:** *What unique linguistic patterns (i.e., topics) in the posts lend more nuanced insight into the differing contexts in which these sexual risk experiences unfold?*

To answer these research questions, machine learning models were trained based on manually labeled data ($N = 8,271$) to first identify disclosures of online sexual risk experiences from other types of posts, then to classify these posts based on the expressed level of consent (i.e., consensual, non-consensual, sexual abuse) and by relationship type (i.e., stranger, friend/dating, family). The classifiers were then used to machine label a larger corpus of posts ($N = 45,955$) for further analysis. Between-group differences were examined through a Chi-square (χ^2) analysis of the larger dataset. Our analysis revealed that the sexting experiences of youth vary significantly based on the relationship type with the other person involved. We then leveraged topic modeling and our own qualitative insights. Overall, we found that there were beneficial reasons for teens to engage in consensual sexting, such as when exploring their sexuality. However, in other cases, even consensual online sexual experiences were often due to underlying mental health conditions. This study makes several important contributions to the fields of Human-Computer Interaction (HCI) and the literature on adolescents' online safety:

- We accurately classified the levels of consent and relationship types within these disclosures (average AUC=90), demonstrating the importance of the underlying architecture of machine learning models to achieve accurate classifications.

- We uncovered key patterns related to teens' online sexting behaviors based on the level of consent and relationship types, showing that teens experienced more consensual sexting with friends/dating partners, non-consensual sexting with strangers, and sexual abuse committed by family members.
- We highlighted the nuances unpacked from teens' disclosure that went beyond the explicit consent statement to understand the underlying factors (e.g., mental health issues) that surround and may undermine consent.

Our results demonstrate the complexities around technology-mediated consent, and online sexual risks, and how these experiences cannot be studied in isolation from the mental health and well-being of youth.

3.2 Background

The literature on adolescent online safety has primarily focused on protecting adolescents from exposure to possible threats (360). As such, research on adolescents' sexting behaviors is often motivated by the potential risks, adverse consequences, and legal considerations (100; 169; 29; 37; 348). In this section, 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.

3.2.1 Adolescent Online Safety and Sexual Risk

Empirical studies directly involving youth tend to highlight the adverse outcomes associated with the online sexual risk behavior of minors. For example, Gamez et al. (124) found that adolescents who have experienced sexting in the past will most likely experience a significant increase in the number of sexual solicitations over the next year. Galvete et al. (62) found that adolescents who were solicited by adults online engaged in more sexting experiences with others. Other studies have highlighted the negative outcomes and life consequences of online sexual risk behaviors, including mental health problems, teen pregnancy, sexually transmitted diseases, and drug and alcohol abuse (29; 71; 52; 338). Another major line of research suggests that online sexual experiences with adult strangers (i.e., sexual predators) can entail the most serious adverse consequences, such as sexual solicitations (234; 345; 346). Overall, these studies emphasize the heightened risk associated with teens engaging in sexual behaviors online, which in turn highlights the need for effective risk mitigation and prevention strategies to protect youth online. Yet, scholars (281; 245) have recently started to push back on the intense "moral panic" around the technology-mediated risks posed to youth, suggesting we take a more child- and teen-centric approach to studying risk-related online phenomena. For instance, Gewirtz-Meydan et al. (130) surveyed youth to understand their perceptions and attitudes around sexting and found that adolescents who engaged in sexting did not view it as a crime. Instead, Razi et al. (288) recently found that teens viewed sexting as a normal progression of their romantic relationships and garnered some benefit from these experiences. As such, researchers have begun to advocate for the importance of acknowledging both positive and negative developmental outcomes associated with adolescent sexting with a focus on educating adolescents about safe sexting practices (243; 271). Therefore, addressing the perception gap between overly risk-focused research and adolescents' personal experiences regarding their online sexual encounters necessitates a deeper and more nuanced examination and teen-centered understanding of these experiences.

3.2.2 Computational Approaches to Detecting Sexual Risks

In recent years, researchers have started to apply deep learning methodologies, such as Convolutional Neural Network (CNN), to detect sexual risks in social media data (170; 204; 373; 81; 173); with promising results. For instance, Chowdhury et al. (81) applied various deep learning models to identify disclosures of sexual harassment using public Twitter posts and achieved 96% accuracy. Hassan et al. (150) proposed data-driven supervised learning for identifying sexual violence reports from the #MeToo movement on social media to examine these types of disclosures more deeply. These researchers detected whether the posts included sexual violence, distinguished among different types of sexual violence (e.g., Unwanted Sexual Contact, Non-contact Unwanted Sexual Experiences, Sexual Violence, Completed or Attempted Forced Penetration, Alcohol or Drug Facilitated Penetration, Forced Acts, Alcohol or Drug Facilitated Acts), and identified the relationship between the perpetrator and the victim(s) (e.g., Intimate Partner, Family Member, Person in Position of Power/Authority/Trust, Friend or Acquaintance, Stranger) (150). Their best F1 score reported for detecting sexual violence was 80%, with 58% for specific type of sexual violence and 62% for the perpetrator-victim relationship. A recommendation from this study was to address the lower accuracy of the classifiers by adopting a deep learning approach.

We build upon and address the limitations of these related works in several ways. First, prior work was not focused specifically on detecting the online sexual risk disclosures of adolescents, who are a particularly vulnerable class of internet users. In contrast, our work focuses specifically on teens (ages 13–17) and on their first-hand online sexual risk experiences. Second, we contextualized our classifiers based on the level of consent expressed in the post (i.e., consensual sexting, non-consensual sexting, sexual abuse) and relationship type (i.e., Stranger, Friend/Dating Partner, Family), obtaining accuracy levels that were better or on par with past studies. Third, we move beyond the risk classification problem to further unpack statistical differences and qualitative insights from the digital trace data of teens who sought advice or support regarding their online sexual risk experiences. In doing so, we take a more holistic approach to detecting and understanding the myriad online sexual risk experiences encountered by modern-day teens.

3.3 Methods

Our main goal was to unpack adolescents' online sexual risk experiences by deeply analyzing their posts disclosing these situations. To accomplish this goal, we first describe our dataset, scoping process, and data annotations for ground truth. Then, we describe how we addressed each of our research questions.

3.3.1 Dataset

3.3.1.1 An Online Peer Support Platform for Youth and Young Adults

We licensed a dataset from an online peer support platform that caters to youth and young adult users, who are interested in discussing topics related to mental health, relationships, sexuality, religion, and more. We chose to anonymize the name of this platform to protect the identities of the youth whose data we analyzed. On this platform, youth post pseudonymously (i.e., by username rather than by real name (216)), and we took care to remove any personally identifiable information from quotations shared in this paper. The dataset originally contained around five million posts and 15 million comments made by approximately 400,000 users. The posts' time frame ranged from 2011 to 2017. Approximately 70% of the platform users were between the ages of 15–24. Although the dataset did not contain any information about nationality, most platform users were English speakers. Our Institutional Review Board (IRB) deemed this study to be 'non-human subjects' research because we analyzed a dataset without personally identifiable information (e.g., usernames). For the protection of users' privacy, the quotes included in this paper were paraphrased or slightly altered (e.g., adding abbreviations and introducing false details that do not affect the context (56)) to make sure the quotes could not be linked to specific people.

3.3.1.2 Data Scoping and Relevancy Coding

The dataset was scoped and annotated as part of our prior published work (anonymized for review). Therefore, we will provide the necessary details needed below for review and reference our prior work upon publication. In order to scale down the five million posts into a practical number of posts for data annotation, we took the following steps. First, we filtered the posts based on user-labeled categories provided by the platform when a user created a new post. The relevant categories included sex, relationships, friends, family, ask girls, and ask guys. We determined the most relevant categories based on a manual inspection of the data. Second, we filtered the posts to include only users who were between the ages of 13 and 17 based on their profile information that was provided with the dataset. The resulting dataset contained 54,226 posts. Third, we filtered the posts to include those that contained both sexual and technology-related words. To do this, we created a lexicon of popular sexual jargon used by adolescents (Team) combined with technology-oriented terms, such as the names of popular social media platforms. These search terms were also supplemented by additional keywords extracted after a manual review of five thousands posts. The keywords used were grouped conceptually into "social media platform," "online," and "sexual" categories and listed in Table 3.1. Then, a SQL query was written based on these keywords to filter the 54,226 posts, resulting in a set of 8,271 posts made by 6,351 adolescents about their online sexual risk experiences.

Types	**Keywords**
Social Media Platforms	Facebook, Instagram, Tinder, Bumble, Grinder, Snapchat, Craigslist, Skype, Hinge, Whatsapp, Kik, Discord, Messenger, Omegle, Vimeo, Vine, Tumblr, Myspace, 4chan, Reddi, forum, blog, Facetime, ft
Online/Technology Terms	video chat, message, dm, sent, send, pm, online, meet on, met on, webcam, gaming, cyber, blackmail, internet, AMOSC, f2f, LMIRL
Sexual Jargon	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, non-consensual, 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.

Table 3.1: Scoping Search Terms. The acronyms' definitions are listed in Appendix A.

Next, we reviewed the 8,271 posts for relevancy. Posts were deemed relevant if they described some kind of sexual experience that involved an online component. Posts were divided among five annotators, and each post was reviewed by two coders. The raters showed a substantial agreement (Cohen's kappa = 0.71). A consensus was formed among all five coders to resolve conflicts. The resulting dataset contained 4,180 (51%) disclosures about online sexual experiences and 4,091 (49%) posts that did not meet this criterion. These labels (online sexual disclosure/not online sexual disclosure) were then used as ground truth labels for addressing RQ1.

3.3.1.3 Ground Truth Annotations

The 4,180 relevant posts disclosing teens' online sexual risk experiences were further annotated based on 1) *level of consent* and 2) the *relationship type* between the teen and the individual in which they described sharing the sexual experience. Two independent annotators coded each post and reached substantial (Cohen's kappa ¿ 0.70) to complete (1.00) agreement. We describe these dimensions and codes in more detail below.

3.3.1.3.1 Levels of Consent.

While consent is a complex concept when dealing with minors, who are by legal definition under the age of consent (186), our work acknowledges the importance of taking into account the agency and first-person perspective of teens when disclosing their personal sexual experiences. Further, prior work on adolescent online risk behavior has shown that teens' online risk experiences vary significantly based on whether they considered themselves victims, willing participants, or initiators of a given risk experience (362; 361). Through a grounded analysis of the data, we derived the following distinct levels of consent:

- *Consensual Sexting:* Posts where teens explicitly stated that they pursued or willingly participated in an online sexual exchange with another person.
- *Non-consensual Sexting:* Posts where teens explicitly stated that the online sexual exchange was unwanted, unwelcomed, or unsolicited.
- *Sexual Abuse:* Posts that disclosed non-consensual online sexual exchanges evolved into a physical sexual interaction in real life (e.g., statutory rape).
- *Consent Status Unknown:* Posts where the level of consent was not expressly specified or discussed in an interpretable way, or posts where consent as a concept was unknown.

As noted above, we used this definition of consent, where teens had to explicitly state a willingness to engage in a sexual exchange for it to be coded as 'consensual.' Our rationale for this decision was that interpreting implied consent from a single post is problematic from both an ethical and legal standpoint (41). Thus, posts had to be clear that the sexting behavior was either initiated by the teen or done willingly (without undue coercion). For example:

> *"Anyone sext? I've been sexting for a year and it's like an addiction to flirt with people online once I'm feeling bored" –Female 14 years old.*

Regarding non-consensual sexting, we found that teens were often fairly direct about their lack of consent, which aided in coding these instances:

> *"Why do guys just send nudes? All I said was hey and then you sent me a dick pic without asking! What makes you think I would want that? I'm not a hoe, and sorry I'm not gonna entertain you" –Female 16 years old.*

Sexual abuse involved non-consensual sexting that resulted in an offline sexual encounter with the teen. While we acknowledge that all non-consensual sexting can be viewed as a form of sexual abuse (82), these situations were different in that they were *physically* harmful to youth and would potentially rise to the level of mandated child abuse reporting (228); therefore, a separate category for these types of sexual risk experiences was warranted. As an example, many teens disclosed their personal stories about being victims of rape:

> *"My relationship with my boyfriend started when i was 13. after about a week I sent him videos... I was happy to make him feel good. After that he started to touch me in class... Later on he raped me. I didnt want it happening again. He told me I was worthless girl with only a body and a slut. I became extremely suicidal I took pills and I tried to cut" – Female 14 years old.*

We coded a total of 1,136 posts as consensual sexting, 705 as non-consensual sexting, 243 as sexual abuse, and 2,043 as not applicable.

3.3.1.3.2 Relationship Types.

Relationship type is another important aspect of teens' online sexual experiences, as the assessment of the sexual riskiness attached to these experiences may vary based on the relationship between youth and with whom they sext (341; 288). The relationships between teens and others involved in the online sexual risk experience disclosure posts were:

- *Strangers:* Posts that describe experiences between an adolescent and a stranger.
- *Dating Partners:* Posts that describe experiences between an adolescent and a dating partner in a romantic relationship.
- *Friends:* Posts that describe experiences between an adolescent and a friend or acquaintance.
- *Family Members:* Posts that describe experiences between an adolescent and a family member.
- *Not Applicable:* Posts that were ambiguous as to the relationship between the teen and the individual or did not specify.

Quotations presented in section 4.4 are conceptually grouped by the relationship types described in the posts to illustrate examples of how we coded based on relationship type. While these categories were relatively straightforward to code for, during our preliminary analysis of the data set we noticed enough similar patterns between the "Dating" and "Friend" categories that it was often difficult to ascertain the differences between the two. Further examination of the literature revealed that past studies also found the boundary between friendship and romantic relationships was often blurred in adolescence (84; 86). For instance, as in the following quotation, teens often talked about romantic feelings towards someone described as a close friend.

> *"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?" –Female 14 years old.*

Further, when conducting an initial analysis for RQ2, the χ^2 test of independence found no significant differences between the 'Dating' and 'Friend' categories [χ^2 ($df = 4, N = 2084) = 210.36, p = 0.87$]. For these reasons, we decided to combine the two into one Friend/Dating category for the analysis presented in this paper. Therefore, we coded 2,084 disclosures about online sexual experiences that occurred with strangers, 841 within friend/dating relationships, and 80 with family members. These codes were used to train a relationships classifier model (as part of RQ1) to machine label a larger dataset for further analysis in RQs 2 & 3, while the posts that were coded as not applicable were treated as missing values for the machine learning algorithms.

3.3.2 Classifying Online Sexual Risk Disclosure Posts

In the following sections, we explain our data pre-processing and the supervised machine learning approach for answering RQ1.

3.3.2.1 Data Pre-processing and Models

Multiple steps were performed to pre-process the datasets before running the models. First, any post with one or two words was removed since context cannot be extracted from two words. Then, the posts were converted to all lowercase. A preliminary exploration was done for the classification framework with stopwords and stemming, and we found no noticeable differences in the accuracy of the classifiers. Therefore, we opted to keep the stopwords and the original form of the words to preserve how adolescents express themselves. The next step was using a Python library called Keras Tokenizer to convert the posts into tokens that can be fed to the models. Both interpretable models (SVM, Random Forest, and Logistic regression) and deep learning models were preliminarily explored, and the initial results yielded from training the models showed that deep learning models significantly outperformed the interpretable models; therefore, we opted to move forward using deep learning models. The first step was to train and optimize the classification models using the manually labeled dataset. In this work, we applied deep learning models for predicting the following:

- *Online Sexual Risk Disclosures/Not Online Sexual Risk Disclosures*: Binary classification models used to predict whether or not a post contained an online sexual risk disclosure.
- *Level of Consent*: Multi-class classification models were used to predict the types of sexual risk experiences based on the levels of consent, which were consensual sexting, non-consensual sexting, and sexual abuse.
- *Relationship Types*: Multi-class classification models were used to predict the relationship types, which included stranger, friend/dating, and family.

Deep learning models are known to decrease the false rates for text classification (43). Based on this fact, the Long Short-Term Memory (LSTM) model and the Convolutional Neural Network (CNN) model have been widely applied for text classification. Recently, CNN and LSTM have been used for Natural Language Processing of small text classification (272; 181). Therefore, the performance of CNN and LSTM across the three text classification tasks was used, explored, and compared in this study. A 5-fold cross-validation was conducted along with a random search to tune the hyperparameters for each model. In each fold, 80% of the data was used for training (out of this 80% of the data, 10% was used as the validation set) and 20% of the data was used for testing. The next section will discuss in more detail the evaluation matrices we applied to compare the models' performance.

3.3.2.2 Evaluation

Since we applied the 5-fold cross validation, the average accuracy of the models, the standard deviation of the accuracy, the class-specific precision and recall, the F1-measure, and the area under the receiver operating characteristic (ROC) curve (AUC) were used to evaluate our models using the test sets. The accuracy and F1 scores provide general insight into the performance of the models; therefore, only AUC was analyzed in this study since it can provide more detailed insights. We report the performance metrics of our classifiers in section 4.1.

3.3.2.3 Machine Labeling

After training and evaluating the models based on the manually annotated ground truth data, the trained classifiers with the best accuracy performances were then used to machine label the rest of the dataset beyond the manually annotated data. We labeled the entire dataset ($N = 45,955$) based on the classifiers for online sexual disclosures/not online sexual disclosures, levels of consent, and relationship types, which identified a total of ($N = 25,808$) posts that contained an online sexual risk disclosure made by teens. Machine labeling the entire dataset increased our power to detect significant differences for the χ^2 tests in RQ2, and the larger number of posts gave us the ability to detect more nuanced topics (RQ3) and better understand adolescents' online sexual disclosures across the dimensions of levels of consent and the relationship types.

3.3.3 Examining Between-Group Differences

To examine whether there were differences between the three types of sexual risk experiences and the three relationship types (RQ2), we performed a χ^2 test of independence, which are between-group (rather than within-group) tests applied when there are two or more nominal variables, each with two or more possible values (316). The standardized residuals are calculated by dividing the product of subtracting expected from observed values by the square root of the expected value as an estimate of the raw residual's standard deviation (316). The standardized residuals were used in this study to show the significant differences between the relationship type and the types of sexual risk experiences. The χ^2 test was conducted for the combined dataset ($N = 27,892$) comprising the manually annotated and machine labeled posts.

3.3.4 Topic Modeling Approach

For RQ3, we leveraged topic modeling (137) to further unpack teens' online sexual disclosures across the differing levels of consent and relationship types. Topic modeling, which has become a popular approach in the HCI literature (138), is a useful unsupervised approach to identify topics in teens' online sexual disclosures based on a textual analysis of these documents. Similar to prior works (134; 35), we complemented the topics extracted by the algorithms with our own qualitative interpretations of the data. To do this, we analyzed the top 15 words contributing to the topic and then read through the top 50 ranked posts (with the highest probability for each topic) through an iterative and inductive qualitative content analysis (158) approach to further interpret contextual details contained within the disclosures. We semantically labeled our topics to assign high level descriptors to them based on our understanding of the top 15 keywords and the qualitative interpretation of the top 50 posts.

For creating the topics, the posts were cleaned based on the following steps: 1) removing stopwords that did not add any semantic value, 2) stemming, 3) tokenizing. After the normalization step, the topic model was iteratively run and kept removing words that were not specific enough or meaningful to the analysis, such as "the," "and," "or," and common pronouns. We then proceeded to run the model on the rest of categories. We reported the average coherence score for each number of topics to identify the number of topics that would provide succinct cohesion for a particular category of posts as shown in Table 3.3.

Two different topic modeling approaches were experimented with for this study: 1) Latent Dirichlet Allocation (LDA) (47) and 2) Dirichlet Mixture Model (DMM), which is specifically designed for overcoming the sparse and high-dimensional problem of clustering short texts (379). To choose the best one, we experimented with the two selected approaches by applying them across categories with different numbers of posts: Stranger and Consensual Sexting, Friend/Dating and Consensual Sexting, and Family and Sexual Abuse. To evaluate the quality of the yielded topics, we used coherence score, which measures the degree of semantic similarity between the top keywords of the topic (231). We ran the two models with different starting numbers of topics (from 2 to 15) to compare the performance of these models based on the average coherence scores. LDA performed poorly compared to DMM across the selected categories, especially the category of family and sexual abuse, which had the smallest number of posts (comparing with the three selected categories). The best average coherence score for LDA was for the stranger and consensual sexting category (the category with the largest number of posts), but this was less than DMM (a larger coherence score means the topics are more coherent). Overall, the DMM showed the best average coherence scores (avg.coherence: -79.39) and yielded more semantically interpretable topics in comparison with the LDA model (avg.coherence: -98.02) as shown in Table 3.2. Therefore, we proceeded with the DMM algorithm. For the rest of the categories, we followed the same procedure by running the DMM algorithm with different numbers of topics (from 2 to 15) to determine the best number of topics for that category based on the best average coherence score as listed in Table 3.3.

Model	**Topic 1**	**Topic 2**	**Topic 3**
LDA	school, thought, happen, everything, rachel, wrong, anyway, problem, another, almost, kik, skype, send, among.	listen, dear, okay, world, happy, left, haven, guy, story, have, know, couple, comment, best, what.	sexual, weird, scare, gonna, kinda, worth, sound, internet, problem, video, conversation, comment, video, start, hard.
DMM	bisexual, crush, question, single, lesbian, skype, attract, account, roleplay, advice, kik, gay, anyone, flirt, send.	depress, sext, addict, attention, want, suicide, random, wonder, feel, cut, change, understand, better, help, need.	older, sext, snapchat, ugly, talk, pretend, instagram, account, restrict, creep, flirt, hot, age, men, lie.

Table 3.2: The top 15 words in topics discovered by LDA and DMM.

Category	**K=2**	**K=3**	**K=4**	**K=5**
Consensual Sexting & Stranger	-99.56	**-79.39**	-108.60	-89.32
Non-consensual Sexting & Stranger	**-133.65**	-199.02	-167.87	-168.98
Consensual Sexting & Friend/Dating	**-57.43**	-84.12	-102.34	-190.45
Non-consensual Sexting & Friend/Dating	**-65.23**	-87.21	-91.62	-88.43
Sexual Abuse & Family	**-43.12**	-66.80	-79.32	-94.21

Table 3.3: Average coherence score on the K number of topics for each category. A larger coherence score means the topics are more coherent.

For each topic, we then conducted a qualitative analysis to gain further insights. We did this by first analyzing the top 15 most probable keywords for a given topic, then the first author did a qualitative reading to analyze the top 50 representative posts with the highest probability for each topic. We used this qualitative understanding of the posts and top keywords to answer RQ3. Section 3.4.4 presents the topic models for the dimensions that yielded strong statistical significance in our RQ2 analyses and includes illustration quotes to contextualize each topic. For smaller categories (i.e., sexual abuse by strangers and friends/dating), where there were low counts that affected the topic modeling to not yield strong coherence scores, we provided our own qualitative insights to characterize the posts in these categories.

3.4 Results

In this section, we begin by answering RQ1 and demonstrating how CNN outperformed LSTM for the three classifiers we built for online sexual disclosures, levels of consent, and relationship types. Then, we present the significant different patterns of associations between the levels of consent and relationship types to answer RQ2. Finally, we explore the different topics yielded from the significant associations based on the between group analysis to unpack the different levels of consent and relationship types.

3.4.1 Identifying Online Sexual Posts (RQ1)

The first step to evaluate the overall accuracy performance of the two deep learning models was calculating the baseline accuracy, which was 51% for the online sexual disclosures/not online sexual disclosures classifier, 36% for levels of consent classifier, and 37% for the relationship types classifier. Table 3.4 reports the accuracy metrics for the three classifiers; it clearly shows that the overall performance is better than the baseline accuracy for the three classifiers. For the online sexual posts classifier, the CNN model performed better than LSTM ($AUC = 0.93$) in identifying online sexual posts (refer to Figure 3.4 for ROC curve of online sexual classifier and to Table 3.4). For class-specific performances, we observed that CNN yielded a higher precision and recall results for both classes than LSTM, which confirms the accurate predictions of CNN for each class. The online sexual classifier labeled 56% (25,808 out of 45,955) as online sexual disclosures and 44% (20,147 out of 45,955) as not online sexual disclosures.

The CNN also outperformed LSTM ($AUC = 0.84$) in predicting types of levels of consent (consensual sexting, non-consensual sexting, sexual abuse) as shown in Table 3.4. Figure 3.4 shows the ROC curve of the levels of consent classifier. Out of the 25,808 online sexual disclosures, we obtained 60% ($N = 16,616$) as consensual sexting, 28% ($N = 7,933$) as non-consensual sexting, and 12% ($N = 3343$) as sexual abuse.

Finally, in identifying relationship types in the posts (stranger, friend/dating, and family), CNN also performed better than LSTM (AUC = 0.91) as shown in table 3.4. The ROC curve of the relationship types classifier is shown in Figure 3.4. For the relationship classifier, we obtained 59% ($N = 16,355$) labeled as stranger, 34% ($N = 9,466$) labeled as friend/dating, and 7% ($N = 2071$) as family. The analysis for RQ2

Models	**Target**	**Classes**	**Avg.Acc**	**SD**	**Prec**	**Recall**	**F1**	**AUC**
CNN	Online Sexual Disclosures	Online sexual	0.91	0.47	0.94	0.92	0.93	**0.93**
		Not online sexual			0.88	0.90	0.89	
LSTM		Online sexual	0.82	0.69	0.83	0.91	0.87	0.85
		Not online sexual			0.88	0.72	0.79	
CNN	Levels of Consent	Consensual Sexting	0.80	1.02	0.85	0.89	0.87	**0.84**
		Non-consensual Sexting			0.60	0.62	0.61	
		Sexual Abuse			0.81	0.83	0.81	
LSTM		Consensual Sexting	0.72	1.91	0.78	0.74	0.75	0.75
		Non-consensual Sexting			0.43	0.38	0.40	
		Sexual Abuse			0.50	0.71	0.58	
CNN	Relationship Types	Stranger	0.88	1.79	0.91	0.95	0.93	**0.91**
		Friend/Dating			0.92	0.85	0.88	
		Family			0.79	0.80	0.79	
LSTM		Stranger	0.80	1.42	0.83	0.79	0.81	0.85
		Friend/Dating			0.76	0.81	0.78	
		Family			0.72	0.80	0.75	

Table 3.4: Metrics of deep learning classifiers in *k*-fold (*k*=5) cross-validation.

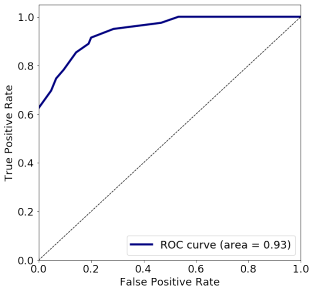
and RQ3 will proceed with a combined dataset of the manually annotated and the machine labeled posts ($N = 27,892$). Figure 3.5 (A) shows the distribution of the combined dataset across the types of levels of consent and relationship types.

3.4.2 Characteristics of the Combined Dataset (Machine and Manually Labeled)

The 27,892 relevant online sexual disclosures were written by teens who were between 13 and 17 years old, with an average age of 15 years old (at the time of posting). Fifteen-year-old teenagers were represented the largest percentage of posts (27%), followed by users who were 16 (20%), 14 (19%), 17 (18%), and 13 (16%). Most of the posts were written by female users (78%), followed by male users (12%) and non-binary or unspecified gender users (10%). These teens had been active in the platform for an average of 7.5 months from the date they first posted. Based on the teens' posts, 36% mentioned using other social media platforms, such as Kik (39%), Skype (15%), Snapchat (12%), Facebook (12%), Instagram (9%), a peer support platform (8%), Tumblr (3%), and Omegle (2%).

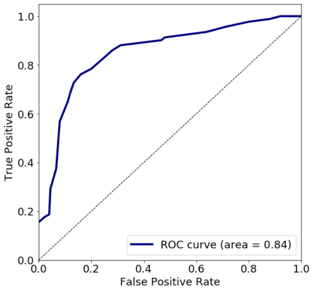
3.4.3 Sexual Risk Experiences Significantly Differ Based on Level of Consent and Relationship Type (RQ2)

A χ^2 test indicated a significant association between relationship type and levels of consent [χ^2 ($df = 4$, $N = 27,892) = 3357.4, p < 0.001$, Cramer's $V = 0.245$]. Post hoc testing revealed all three relationship types were significantly different from each other in terms of the level of consent. Specifically, after p-values were adjusted using Bonferroni correction, there were significant differences between stranger and friend/dating ($p < 0.001$), between stranger and family ($p < 0.001$), and between friends/dating and family



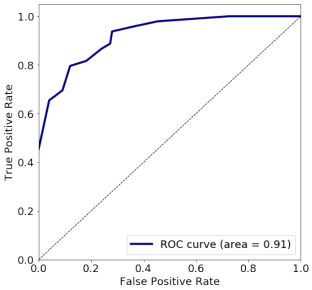
.3

Figure 3.1: Online Sexual Disclosures



.3

Figure 3.2: Levels of Consent



.3

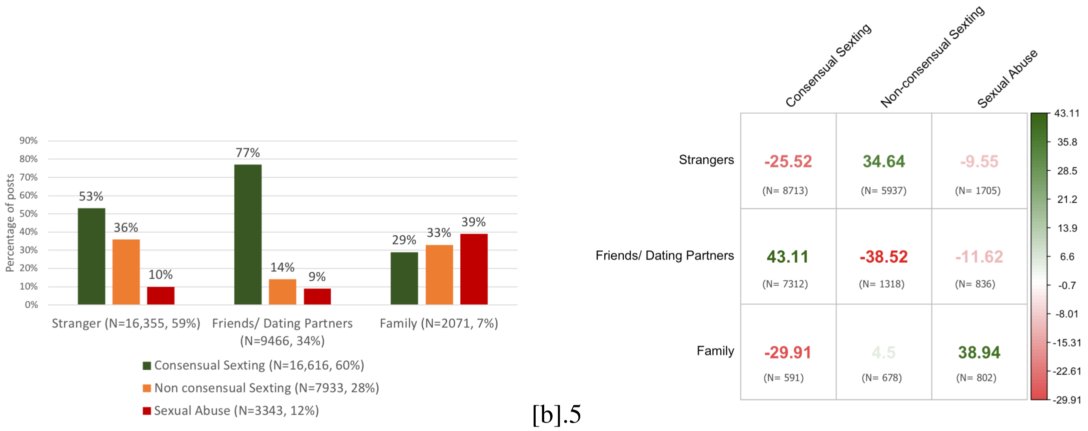
Figure 3.3: Relationship Types

Figure 3.4: ROC Curves of the classifiers with the best accuracy

.

($p < 0.001$). As illustrated in Figure 3.5, when comparing proportions of level of consent disclosures for each relationship type, for consensual sexting, the friend/dating category had the highest relative proportion (i.e., 77% of experiences involving friend/dating relationships were consensual sexting), followed by stranger (53%) and family (29%). For non-consensual sexting, stranger had the highest relative proportion (36% of stranger-related experiences were non-consensual sexting), followed by family (33%) and friend/dating (14%). For sexual abuse, the family category had the highest relative proportion (39% of adolescents' online sexual experiences with family were sexual abuse), stranger (10%) and friend/dating (9%).

Next, we used the standardized residuals to more closely examine associations between each relationship type and levels of consent. Each pairwise comparison showed a significant association (shown in Figure 3.5, all adjusted p-values < 0.001 using Bonferroni correction). First, for experiences involving strangers, there



[b].5

[b].5

Figure 3.5: (A) Distribution of levels of consent and relationship types of the combined dataset ($N = 27,892$). The bars are standardized to 100% for each category. (B) Correlation matrix of Pearson's standardized residuals between the levels of consent and relationship types.

was a strong negative association between strangers and consensual sexting and a strong positive association between strangers and non-consensual sexting. These diverging associations suggest that when teens post about sexting with strangers, discussions are more likely to involve non-consensual interactions. In contrast, for experiences involving friends/dating partners, we observed the opposite pattern for consensual and non-consensual sexting. Specifically, there was a strong positive association between friend/dating and consensual sexting and a strong negative association between friend/dating and non-consensual sexting, suggesting teens are more likely to have or discuss consensual sexting as a positive experience with friends or dating partners, while non-consensual sexting was less likely in these relationships. Both stranger and friend/dating relationships had a relatively weak negative association with sexual abuse. In contrast, for experiences involving family, there was a strong positive association between family and sexual abuse, suggesting discussions about family on the peer support platform studied may be relatively more likely to involve sexual abuse compared to discussions about other relationships. Additionally, there was a strong negative association between family and consensual sexting, and a weak positive association between family and non-consensual sexting. These findings provide interesting insights and allow us to unpack these strong associations to examine the nature of these discussions, which we explore in the following section.

3.4.4 Unpacking the Online Sexual Risk Experiences of Teens (RQ3)

In this section, we discuss the topic modeling results and qualitative interpretations to unpack the differences between the levels of consent and relationship types reflected in the posts from the combined dataset (manually and machine labeled data). The high level descriptor of the resulting topics along with the top 15 keywords of adolescents' online sexual disclosures across levels of consent and relationship types are displayed in Table 3.5.

3.4.4.1 Sexting with Strangers.

In this section, we unpack the emerged topics from youths' posts about their consensual and non-consensual sexting experiences with strangers online.

3.4.4.1.1 LGBTQ+ and sexuality ($N = 4,443$, 51%).

The largest percentage of posts regarding consensual sexting with strangers were posts made by teens who identified as part of the Lesbian, Gay, Bisexual, Trans-gendered, and other sexual minority (LGBTQ+) community. These adolescents referenced using online private messaging platforms, predominantly Kik, to connect with and solicit attention from others within the LGBTQ+ community (e.g., crush, single, attract, flirt, roleplay). For example, teens often asked to connect with others of a specific sexual orientation to engage in sexual interactions ranging from simple flirtations to sexual 'hook-ups.' A common pattern among these posts was that these teens often did not specify their gender in their profiles, likely as a way to signal their non-conformance to gender norms.

> *"I want a flirting buddy .. any takers? bisexual or lesbian .. only girls. send your kik." – Unspecified gender 15 years old.*

In many cases, these teens were using a platform that meant for mental health support as a dating app by posting requests to connect with other teens for sexual interplay. While most of the posts were from LGBTQ+ teens looking for partners to sext, in other cases, teens sought to explore their sexual identities through sexual roleplaying with strangers. It was less clear in these cases whether the teens identified as part of the LGBTQ+ community or simply wanted to explore different sexual orientations. For example, a 14-year-old (unspecified gender) asked the following question in a post:

> *"anyone wanna do gay roleplay with me on skype? Ill still love you if you dont ??" –Unspecified gender 14 years old.*

In other cases, teens sought advice from strangers about their sexuality and/or sexual orientation.

> *"Bisexual & Gay Males I need your advice- I'm 16 and questioning my sexuality- I have a question: Is it strange that I am attracted to: the idea of online sexual relationships with guys, the idea of giving and receiving oral sex from a guy, BUT I am not okay with other forms of sex, and I don't know how id feel about kissing." –Unspecified gender 16 years old.*

Relationship Type	**Topics**	**Top 15 Keywords**
With Strangers ($N = 16,355,59\%$**)**		
Consensual Sexting ($N = 8713,53\%$) (avg. coherence: -79.39)	LGBTQ+ and sexuality ($N = 4443,51\%$)	bisexual, crush, question, single, lesbian, skype, attract, account, roleplay, advice, kik, gay, anyone, flirt, send.
	Mental health motivated sexting ($N = 2962,34\%$)	depress, sext, addict, attention, want, suicide, random, wonder, feel, cut, change, understand, better, help, need.
	Sexting with/as adults ($N = 1306,15\%$)	older, sext, snapchat, ugly, talk, pretend, instagram, account, restrict, creep, flirt, hot, age, men, lie.
Non-consensual Sexting ($N = 593,36\%$) (avg. coherence: -133.65)	Harassment and sextortion ($N = 3767,62\%$)	threaten, blackmail, stupid, pic, scare, screenshot, videoed, rape, police, snapchat, convince, internet, donot, send, demand.
	Unsolicited Exposure ($N = 2170,38\%$)	guy, pic, send, dick, kik, nude, scare, random, ask, pervert, talk, puke, me, traumatise, get.
With Friends/Dating Partners ($N = 9466,34\%$**)**		
Consensual Sexting ($N = 7312,77\%$) (avg. coherence: -57.43)	Sexting within a close relationship ($N = 4314,59\%$)	flirt, ldr, cheat, sext, skype, sexual, distance, happy, pic, together, trust, fun, hate, boyfriend, friend.
	Mental health motivated sexting with friends/dating partners ($N = 2997,41\%$)	suicide, sex, picture, attention, reason, send, bipolar, better, differ, bad, depress, feel, friend, commit, ask.
Non-consensual Sexting ($N = 1318,14\%$) (avg. coherence: -65.23)	Peer pressure to sext ($N = 843,64\%$)	pressure, trust, bestfriend, threaten, send, ex, scare, kill, fault, sex, reason, photo, dont, friend, confuse.
	Feeling betrayed by trusted others ($N = 579,36\%$)	send, pissed, stupid, upset, block, mad, off, trust, horrible, bullshit, friendship, disturb, hate, nude, friend.
With Family Members ($N = 2071,7\%$**)**		
Sexual Abuse ($N = 802,39\%$) (avg. coherence: -43.12)	Sharing stories of past/current abuse ($N = 425,53\%$)	parent, memory, afraid, stupid, brother, older, sexual, harass, again, dad, night, sudden, father, abuse, disgust.
	Coping with the aftermath of abuse ($N = 376,47\%$)	abuse, dad, touch, scare, anorexia, cut, molest, stori, harm, again, stomach, suicide, threat, depress.

Table 3.5: Emerged topics across levels of consent and relationship types ($N = 27,892$).

While seeking out strangers to become sexual partners had the potential to lead to risky sexual interactions, many of the posts in this topic also illustrated the benefit of being able to connect with a like-minded community to ask questions, get advice, and explore and exercise their sexual orientation in a way that was true to their identities.

3.4.4.1.2 Mental health motivated sexting ($N = 2,962$, 34%).

The second largest topic for consensual sexting with strangers contained posts about sexting that on the surface seemed consensual, but upon further examination appeared to be motivated by underlying mental health conditions, such as depression (e.g., depress, suicide, cut). These teens appeared to use sexting with strangers as a way to cope with low self-esteem, depression, anxiety, and the desire to be loved, accepted, and seen (e.g., feel, better, help, attention, need). These posts often were made as a confession with an undertone of regret and shame, like the following quote from a teen with a mental illness who was not proud of the nudes she sent.

> *"[want to say the truth] I've cut and had depression and I've sent nudes before I'm not proud of. I did send it to people cause I thought it would help me to live more and feel better ..." –Female 15 years old.*

These posts often include teens' experiences with their mental health, including suicidal ideation, self-harm, and loneliness, combined with a desperate need for relief from these issues. Often, these teens knew that their sexting behaviors were unhealthy; they just wanted to feel good and be wanted, similar to the 17-year-old female below who knew she was being used for male pleasure but admitted that it made her feel wanted.

> *"...I know guys just use me to get off but at least someone wants me..." –Female 17 years old.*

Many teens began sexting with strangers as a way to cope with heartbreak or rejection. Then, they became addicted to the attention and could not stop. Unfortunately, engaging in sexting often had the opposite effect by making teens feel worse about themselves instead of better. Many of the posts, like the one below, expressed a deep-seated self hatred due to these experiences.

> *"... he was all I ever wanted and more but now ... he doesn't even talk to me. I changed so much over the year. i started sending people nudes ... but now I'm addicted and getting on. you made me a fucking hoe ... I'm a bisexual depressed fucking bean" –Female 14 years old.*

Although these posts were labeled as consensual because teens willingly sought out and engaged in these sexual interactions, it was clear that these experiences were harmful to them. Later, we discuss the implications of this finding in terms of the complexities around the concept of consent when dealing with minors, particularly those who have mental health challenges, as well as the potential unintended harm of inaccurate risk classifications (i.e., false negatives for identifying consent) when developing risk mitigation interventions.

3.4.4.1.3 Sexting with/as adults ($N = 1,306$, 15%).

The other topic that emerged from consensual sexting with strangers revolved around the concept of age. Age indicators (e.g., older, age, men) were often associated with sexting keywords (e.g., sext, flirt) across different social media platforms (e.g., snapchat, instagram, account) in these sexual disclosures. Teens discussed various platforms' age restriction rules (e.g., talk, restrict, age). In these posts, teens talked about online sexual exchanges with adults or adults pretending to be teens. In some cases, teens pretended to be adults themselves for the purpose of sexting with strangers (e.g., pretend, lie, creep). In many cases, teens appeared motivated to disclose about these sexual risk experiences because they felt "in over their heads" and needed advice on what to do:

> *"An older female on meetme wants me to sext with her not knowing I'm a lot younger then her ... I had created my account to say I'm older than what I am though I'm only 15 and I know meetme has strict rules but I do not know what to do I asked her if she has kik and of course she said no but I don't feel comfortable at all" –Male 15 years old.*

Many teens expressed a preference to interact sexually with older strangers because they treated them better/different than people who were in their age. In these posts, teens often were attempting to develop caring relationships with people who were older than them, rather than engage in sexting for the act itself. In some cases, their feelings were reciprocated, but in other situations, like the one below, the outcome was unclear:

> *"I'm the kid @ school, who's considered to be the 'ugly girl.' I've always been attracted to older men cuz they don't see me as ugly. They see me as beautiful, sexy, etc... I started a conversation with a man. one of the first things he asked me was can we trade nudes... after a while I gave him a photo of my butt, in a cheeky bodysuit. I truly care about him, and I hope that he truly cares about me too." –Female 16 years old*

Overall, the power differential stemming from the age difference between teens and adults was fairly apparent as the youth often expressed uncertainty or concern and asked others on the platform how best to proceed, as in the following quote from a 17 years old female who sought an advice about the age difference between her and someone who was 10 years older than her.

> *"I have very strong feelings towards a guy I know online. We have sexted and such for a couple months. He's 10 years older than me. He says he likes me. Idk the problem is the age difference. Advice?" –Female 17 years old*

3.4.4.1.4 Harassment and sextortion ($N = 3,767$, **62%**).

Non-consensual sexting was the most prominent and statistically significant pattern we observed when teens engaged with strangers. These posts disclosed risky sexual experiences that occurred on online platforms (e.g., snapchat, internet) that were threatening and/or aggressively charged (e.g., threaten, blackmail, rape, police). Teens described how strangers manipulated and tried to control them (e.g., convince, demand) due to sexual content shared (e.g., pic, screenshot, videoed). Teens expressed fear (e.g., scare) about how to handle these situations and asked others what they should do.

> *"I did something really stupid. I was on a chat site talking to a cute boy when he asked for a bra pic, which I sent, and like an idiot I told him my town and school. Then, everything changed he threatens to take my pic and post it to everyone..." –Female 14 years old.*

As shown above, this type of sextortion often occurred after teens made the mistake of sending a nude or partially nude photo to someone whose affection they were hoping to attract. In more severe cases, teens were then coerced to continue engaging in non-consensual sexting for fear that their past mistakes would be exposed and used against them.

> *"This guy on omegle and he said if I show my tits he will show his dick so I was like okay whatever. I wanted to leave but he said that he videoed me doing it and if I don't do more he will post it on the Internet. So I kept doing and I was scared that he would post them." –Female 16 years*

Strangers used several strategies to harass teens and force them to interact sexually with them. Teens' posts described how strangers blackmailed them by saying they would send pictures or videos of the teen to people they knew with the intention of embarrassing them or getting them in trouble. To do this, the strangers also had to obtain personal information from the teen, such as the names or contact information of their friends or the location of their school and/or home. By the time the person had this information and made their intentions transparent to the teen, it was often too late for the teens to protect themselves from this kind of harassment and abuse. They felt helpless, scared, and ashamed.

3.4.4.1.5 Unsolicited Exposure ($N = 2170$, **38%**).

The other type of non-consensual sexting with strangers involved receiving unsolicited explicit imagery (e.g., pic, send, dick, nude, random, me, get) from random people (e.g., pervert) online (e.g., kik). Teens often expressed negative emotions about these online sexual risk experiences, including fear and disgust (scare, traumatize, puke). These unsolicited messages included nude pictures and pornographic videos. In many cases, teens received these messages ‘out-of-the-blue’ from strangers they had never talked with before. Because of this, younger teens, in particular, found these experiences surprising, disturbing, and even traumatizing.

> *“I’m 13 and someone just sent me nudes I am really traumatised.” –Female 13 years*

Many of the teens said that they blocked the account of the offender, and others also reported the message and the account to the social media platform. It was rare for these unsolicited sexual messages to lead to further sexual exchanges because teens felt blindsided and violated as to why a stranger would expose them to unwanted sexual content in the first place.

3.4.4.1.6 Sexual Abuse by Strangers.

Sexual abuse by strangers yielded weak statistical association in our RQ2 analysis and had a low post count overall. This may be because teens who engaged with others to the point of physical sexual contact no longer considered the other person a stranger. As such, teens’ sexual abuse disclosures in this category mostly described how teens were raped by a stranger that they met online. Teens reached out on the peer support platform to garner support and share their stories to connect with other survivors. As to not trigger the reader, we chose not to include any quotations that depicted these graphic disclosures of rape.

3.4.4.2 Sexting with Friends/Dating.

In this section, we present teens’ consensual sexting, non-consensual sexting, and sexual abuse experiences that involved friends/dating partners online.

3.4.4.2.1 Sexting within a close relationship ($N = 4,314$, 59%).

Over half of the consensual sexting posts in this category involved a romantic partner (e.g., boyfriend, friend). Many of these online sexual experiences (e.g, flirt, sext, sexual, pic) occurred due to long-distance relationships (e.g., ldr, distance) on private messaging platforms (e.g., skype). Unlike sexual disclosures with strangers, teens described deeper relationships (e.g., trust) and more sexual experiences that evoked positive emotions (e.g., happy, fun) within this topic.

> *“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...” –Female 15 years old.*

In long-distance relationships, particularly when teens first met their partners online, sexting was often used as a way to set body expectations and ensure mutual physical attraction before taking the step to meet in person. This meant that sexual exchanges would occur before the two people met in real life, which in some cases made teens more vulnerable, but in other situations, like in the post below, made them feel more secure that they were not wasting their time and affection.

> *"Wanting to send ldr boyfriend who is visiting for first time soon a pic of me in underwear but covering boobs so he can really see what my body looks like and if he likes me for who I am and he's not wasting his time you know? I'm chubby... I think I may feel more relieved about it." –Female 17 years old.*

Teens used consensual sexting as a way to build intimacy and strengthen their relationships. Teens often explained that sexting was a way to make others feel happy or sexually gratified. Interestingly, these exchanges took place not only between partners in a monogamous dating partnership. Close friends or "friends with benefits" also engaged in sexting. However, in these cases, there was more ambivalence and uncertainty in the posts. Teens often worried if they were getting their desired outcome, as in the following quote from a teen who did not feel that sending nudes to his lover made her happy.

> *" I feel as though me sending nudes to this girl I love as a best friend would make her happy and she likes them but I don't feel they make her happy." –Male 16 years old.*

In other cases, sending friends nude pictures backfired when the recipient betrayed the sender's trust by sharing the sexual image with other people without the sender's consent.

> *"I sent nudes to my friend. I know it's stupid but the compliments were so nice and made me not hate myself for awhile, I trusted him. But at school he showed half my grade. I am so embarrassed I cut when I got home and filled the tub with blood. No ones going to look at me the same. I hate him so much, but I hate myself more." –Female 15 years old.*

Situations like the one illustrated above poignantly highlight how expressions of a deep sense of betrayal often were accompanied by disclosures of suicidal ideation and/or self-harm.

3.4.4.2.2 Mental health motivated sexting with friends/dating partners ($N = 2,997$, 41%).

A need for love and acceptance motivated teens to engage in sexting not only with strangers but with friends and dating partners. Disclosures about sexting (e.g., sex, picture) with friends/partners often described mixed emotions (e.g., better, bad, feel), attention-seeking behavior (e.g., attention, ask), and indicators of mental health issues (e.g., suicide, bipolar, depress, commit). In contrast to what we observed with strangers, teens often described engaging in sexting with friends/partners in order to meet the mental health and self-esteem needs of the other person.

> *"My boyfriend makes me feel bad every now and then, always asks to see it when we have cyber sex, when I say soon he now begs and pleads, I like to show in my own time even though he's seen it before .... When he makes me feel bad he blames it on his bipolar and promises the next day to be okay ... What should I do ?" –Female 16 years old.*

If these teens did not want to engage in sexting, they did genuinely want to to help the people who were close to them and were concerned for their safety and well-being. Teens cited suicidal ideation, self-harm, and mental illness as reasons why they felt compelled to sext with friends and/or dating partners.

> *"one of my friends came to me saying how he was going to commit suicide ... then he started talking about pu$$y and sex. Tonight he started asking for pictures and I felt bad for saying no and I sext with him" –Female 15 years old*

This topic sheds more light on the pitfalls of sexual consent, as teens agreed to engage in sexting but did so because they felt guilty and responsible for the mental health and well-being of the people they cared about.

3.4.4.2.3 Peer pressure to sext ($N = 843$, 64%).

Non-consensual sexting disclosures between teens and their friends and/or dating partners shared the same pressure to make the other person happy but with less of a perceived threat of the other person harming themselves as a result of saying "no." In this topic, teens described sexting (e.g., send, sex, photo) experiences associated with more aggressive keywords (e.g., pressur, threaten) with their friends/dating partners (e.g., bestfriend, ex, friend). In most cases, negative emotional keywords appeared in these disclosures (e.g., scare, fault, confuse) as they were non-consensual and unwanted.

> *"my ex boyfriend threatened to leak my nudes that I sent to him when we were dating because I refused to send him more after we broke up" –Female 16 years old.*

In these cases, the sexting encounter had the potential to ruin trust relationships as teens often expressed anger and surprise that their friends would "cross-the-line." As such, teens vented their surprise and frustration by describing how that the experience led them to distance themselves from the other person, who they originally thought could be trusted as a friend.

3.4.4.2.4 Feeling betrayed by trusted others ($N = 579$, 36%).

Many teens went beyond distancing themselves from a friend who tried to engage them in non-consensual sexting to expressing anger, rage, and a sense of betrayal. High levels of negative emotions appeared in teens' posts expressing strong feelings (e.g., pissed, upset, mad, off, horribl, disturb) about their non-consensual sexting experiences (e.g., send, nude) with their friends/dating partners (e.g., friendship, friend). They described these experiences in an unequivocally negative light (e.g., stupid, bullshit, horribl, hate). Some teens had extreme reactions, such as being 'done with life' or hating their situation due to such betrayal from a trusted other:

> *"When you trust your friend, then they ask for nudes. I hate my life!" –Female 16 years old.*

In these posts, adolescents expressed disappointment that their friends did not have more respect for them, which negatively impacted their self-worth and friendship. The unsolicited sexual imagery sent by friends/dating partners made teens feel shocked and violated.

> *"HELP!!!! A friend of mine just sent me nudes over snapchat, I ignored it the first time.. But then he does it again so I block him.. He's a good friend of mine but I'm not into this! And I find it disturbing he would send me something like that.. What should I do??" –Female 14 years old.*

3.4.4.2.5 Sexual abuse by friends/dating partners

While sexual abuse by friends/dating partners was not as common (weak significance in RQ2 and low post count), these posts often disclosed stories of rape by someone the teen knew and/or loved. In many cases, the abuse started with consensual sexting between the known person and the teen, but then went too far when the person forced the teen to have physical sex with them.

3.4.4.3 Sexting with Family Members.

While less common than online sexual disclosures with strangers and friend/dating partners, teens also shared about their sexual experiences with family members. In this case, most sexual interactions with a family member were considered sexual abuse, but there were instances of consensual and non-consensual sexting as described below.

3.4.4.3.1 Consensual sexting with young family members.

Most posts that described a consensual online sexual risk experience with a family member involved similarly aged (i.e., other youth) relatives (e.g., cousins, brother) who wanted to sexually experiment with one another in a non-romantic but sexual way. In many cases, the teens justified the sexual exchange based on their feelings of love or closeness with the other person and how they were sexually aroused by the experience:

> *"My bro is currently doing his studies overseas. We chat on skype almost everynight. Last night, he asked if he could masturbate on cam with me. We are very close to each other since we were little. i confess, i was really turned on watching him." –Female 14 years.*

However, some teens developed romantic feelings towards family members with whom they sexted. In these situations, teens often described a situation where they connected with a cousin online who made them feel loved. Given the familial relationship, these disclosures often were intermingled with a level of confusion about whether the other person reciprocated their feelings or whether a budding romance with a cousin was wrong.

> *"I think I am in love with my cousin. He is 2.5 years older than me and we started talking recently because he popped up on Facebook. We flirt so much and he calls me his baby girl and his princess and tells me that I'm beautiful. but I'm so confused. He's also asked me to toss him off when I see him. Do I love him? Does he love me? " –Female 13 years.*

As shown in the example above, older cousins often engaged their younger cousins in sexting. Young teens may have "consented" to the sexual exchange but were often ambivalent and confused because of it. Some teens also disclosed about their parents finding out about these relationships and punishing them by forbidding the relationship and/or revoking their technology privileges.

3.4.4.3.2 Sharing stories of past/current abuse ($N = 42$, 53%).

When teens disclosed about non-consensual online sexual risk experiences with family members, it was most often sexual abuse. Teens recounted their memories of sexual abuse and rape (e.g., memory, sexual, harass, abuse) committed by an older family member (e.g., parent, brother, older, dad, father) and described the trauma from these experiences (e.g., afraid, stupid, disgust). These experiences often happened regularly and mostly during the night (e.g., again, night). In some cases, other family members were aware of what happened but failed to protect the teen. In many cases, teens disclosed a repetitive pattern of abuse that occurred over months or even years.

> *"my older brother sexually harasses me and once i told my parents then he stopped for 2 months and he kept doing it again" –Female 14 years*

In most cases, these sexual abuse stories involved an interwoven pattern of online and offline sexual abuse that unfolded over time. Similar to the case below, teens were often tricked into sexual exchanges with family members who pretended to be an alternate identity online:

> *"my brother who raped me last summer made a fake facebook account and was messaging me an i didnt know it was him, and we sexted ... i had no idea who it was." –Female 14 years old.*

This type of duplicitous digital sexual abuse was worsened because the family member then has leverage over the teen to further blackmail them into performing other sexual acts against their will. Teens in these situations felt trapped as they were scared of getting exposed for their mistake of sexting in the first place. For this reason, teens often felt like they could not identify their abuser to get help.

3.4.4.3.3 Coping with the aftermath of abuse ($N = 376$, 47%).

Teens used the online peer support platform not only to disclose their abuse but also try to cope with the aftermath. The top keywords that appeared in this topic were indicative of abuse (e.g., abuse, touch, molest) by adult family members (e.g., dad). Mental health indicators associated with these experiences (e.g., anorexia, cut, harm, stomach, suicide, depress) highlight the negative outcomes of enduring childhood sexual abuse.

> *"My mother had a new boyfriend. One night, I was sitting on my bed ... He touched me in ways I never want to be touched again. He hurt me. He threatened me. I was scared. I didn't tell anyone. ... I had enough of feeling worthless. I took about 78 pills, thinking it'd be enough..." –Female 15 years old.*

Teens recounted how these experiences personally affected them (e.g., scare, threat), often mentioning suicidal ideation and self-harm as a response to their abuse. While these experiences were traumatic for the youth, a silver-lining, perhaps, was that the online peer support platform gave teens a place where they could disclose, make sense of, seek support for, and hopefully heal from their abuse. Yet, in many cases, it was unclear from the posts whether the teens were able to get the help they needed to break out of the cycle of abuse and recover.

3.5 Discussion

In this section, we discuss the important implications of our findings. We also present implications for design that move towards support teens' healthy and safe sexual development in online spaces.

3.5.1 Accurately Classifying Sexual Risk Experiences and Relationship Types

For RQ1, our work built upon the body of research that has focused on the automatic identification of online sexual posts using deep learning models. These types of models started to receive more attention in the last three years due to their promising accuracy performance on text classification tasks (76). CNN was found to outperform other models (either traditional or deep learning), on identifying predatory behavior patterns, predatory conversations, and sexual harassment (233; 109; 373). These works presented less accuracy of CNN binary classifiers (up to 0.86) than the average accuracy CNN yielded in this study (0.90). While LSTM architecture is designed for capturing long-term dependency in a sequence of words, CNN focuses mostly on the informative and most useful *n*-grams or keywords from the whole input text (386). The posts in our dataset are considered short-length, which may have contributed to the higher CNN accuracy performance. Although CNN yields state-of-the-art results, it is not one of the most applied machine learning algorithms for detecting sexual risk or its context. Therefore, we recommend using CNN more in sexual risk detection research to provide more evidence that CNN is an appropriate model for this field.

Understanding the differences between CNN and LSTM can help rationalize their different performances. One explanation might be explained by the different architectures of these two models, which illustrates the importance of understanding the semantics of a sentence, something key to deciphering the underlying meaning of postings about teens' sexual experiences. Therefore, CNN was suitable for our dataset since keywords are useful enough to identify the class of each post. In contrast, LSTM's focus on the dependency of words can create noise that make it less useful for terse, casual language used in social media posts. Additionally, sentence length has an impact on the results' accuracy. CNN can take advantage of short sentences while LSTM depends heavily on longer sentences (386). Moreover, stories, social media posts, or conversations are among the most popular examples of the input text types provided for machine learning algorithms to identify sexual risks in general. All these texts can be considered short in length compared to news articles, documents, Wikipedia articles, and other forms of longer text found online. Therefore, future sexual risk detection research could consider examining the performances of these models across the text length and the classification goals (e.g, identifying risky patterns within text) to choose the best classifier to be used based on either keywords in sentences (CNN) or the dependency between the words and/or sentences (LSTM).

3.5.2 Online Sexual Risk Experiences of Teens Vary Based on Relationship

For RQ2, we found statistical significance in teens disclosing proportionally more often about consensual sexting with friends/dating partners, non-consensual sexting with strangers, and sexual abuse with family members. There are several key implications from these findings. First, while prior work has been heavily skewed toward studying teens' sexting behavior with friends and dating partners as a form of peer pressure (347; 44; 147), we found that teens were often consenting participants in these exchanges. This suggests a potential narrative shift in sexual health education towards treating sexting as a normative and developmentally appropriate part of intimate relationships that should be done with safety in mind (51), rather than viewing these experiences as deviant or risky sexual behavior that should be restricted and punished. We also urge scholars to perform longitudinal research to investigate the effect of consensual sexting between friends and dating partners on teens' relationships, relationship skills, and future sexual lives over time.

While sexting with friends/dating partners as a *sexual behavior* may not be unhealthy within teen relationships, from a technical standpoint, the *privacy risk* remains high due to the persistence, replicability, and discoverability (263) of digitally shared sexual information, including nude imagery of minors that could be construed as child pornography. As consensual sexting becomes more prevalent and a common culture among teens, when unpacking the consensual sexting experiences, a concern was warranted regarding the subsequent non-consensual distribution of sexual content that was originally shared consensually between two individuals (known in the literature as "revenge pornography" (300)), which could lead to sextortion and sexual abuse. Therefore, additional research on how best to facilitate *digitally secure* sexual exchanges between consenting adolescents is needed to reduce these types of privacy violations and disentangle privacy risks from sexual risks. For instance, policymakers and legal authorities should find practical ways to lower the burden for teens to report non-consensual distribution of their sexual imagery (i.e., nudes) without fear of legal repercussions for having created and distributed such content. Legal protection holding teens' digital rights to sext with other consenting teens and preventing the unauthorized distribution of this content to others would empower youth to report offenders when unauthorized sharing occurred, which in turn, would prevent sextortion and even subsequent sexual abuse by taking away power from their abusers.

In terms of sexting with strangers, a clear pattern emerged where in teens were frustrated by non-consensual/unsolicited sexual advances from strangers. Our finding contributes to the developing picture of teens' sexting experiences with strangers online, suggesting that teens' struggles might not be related to their consenting to these experiences; rather, the problem might lie with online platforms providing easy ways for for strangers to reach out to teens. This finding supports recent decisions by social media platforms to block strangers from direct messaging minors who are not following them (161). Yet, while such design choices might protect many teens from the potential risks associated with non-consensual sexual interactions with strangers, such changes may also unintentionally hinder the sexual development of LGBTQ+ youth, as we found that sexual exchanges with strangers were sometimes a necessary means for these youth to explore their sexual identity and find like-minded sexual partners. Similar to how prior work has highlighted how content moderation algorithms have unintentionally silenced the voices of those in the LGBTQ+ community (259), we would not recommend implementing blanket policies (e.g., blocking all strangers) that protect heteronormative youth at the expense of youth. Instead, we advocate for providing all youth safer online communities (e.g., verified by age), where they can seek social support and healthy romantic partnerships. As consensual sexting creates the need to raise awareness of the privacy risks involved (e.g., making sure to not show one's face in a nude image (146)), it follows that talking to teens about the potential positive and negative effects of exploring sexuality with strangers requires more nuanced and trauma-informed practices (46).

Importantly, our findings emphasize on how teens leverage online platforms to disclose their sexual abuse experiences. Consistent with prior literature (166), teens in our study often disclosed that sexual abuse was perpetrated by family members. Furthermore, teens described repeated, prolonged, and graphic abuse that occurred in both cyberspace and in the physical world due to a power dynamic they were unable to escape. Consistent with prior research (20), our findings highlight that teens are disclosing their sexual abuse in semi-public virtual spaces, where we might be able to detect these experiences and intervene. Given that teens often do not report their sexual abuse to legal authorities due to shame, fear of retribution, or getting in trouble (313), online platforms could be a first line of defense for teens who may be victims of sexual abuse by delivering just-in-time help resources that either connect them with professionals to report the abuse or educate them on how to take evidence-based measures to protect themselves from sexual violence. These touch-points could prevent such abuse from lasting longer and potentially mitigate the serious long-term impact of extended abuse on teens' mental health. Future research should consider using social media trace data to investigate familial sexual abuse in particular, in line with our findings. These disclosures may provide additional details about the incidents described by the victims themselves and could help supplement gaps in the sexual abuse literature due to a primary reliance on small retrospective reports of sexual abuse obtained from clinical or legal samples (280; 224). Empowered with insights gleaned from these online disclosures, online platforms may find themselves in a better position to support youth in taking an active role in their protection.

3.5.3 A Spectrum of Online Risk Experiences Ranging from Healthy to Harmful

Overall, we found that adolescents experienced a wide range of online sexual encounters, from normatively healthy sexual exploration to sexual abuse. In the sections below, we unpack two complex, intertwined concepts that emerged as important factors when studying these sexual experiences: consent and mental health.

The concept of consent, especially for minors, is inherently complicated and laden with potential misinterpretation, especially when it relates to sexual interactions (42). In our RQ3 results, we unpacked some nuanced examples where explicit consent was given by the teen in their post but the emergent topics within this coded data illustrated that consent was undercut by indicators of mental health problems (e.g., depression, low self-esteem, self-harming behaviors, suicidal ideation). In many cases, teens consented to sext with friends/dating partners due to underlying mental health factors (e.g., threats of self-harm), which raises the question about their ability to give unfettered and well-informed consent for the consequences of their actions. Therefore, we urge future legal and social science research to reconceptualize consent in terms of going beyond the explicit statement of consent to examine the the underlying factors that surround and may undermine it.

In prior work, unwanted sexting (but not overall sexting frequency) has been associated with a higher risk of negative mental health outcomes (343). Other studies have found that sexting is correlated with mental disorders and high-risk behavior (242; 344). The assumption is often that sexting leads to negative mental health outcomes. In our work, however, the topic modeling results uncovered that mental health concerns often acted as an *antecedent* that led teens to engage in sexting as we found teens with mental health issues sought attention by engaging in sexting with others or acquiesced to a sexting request due to concern for the mental health of someone else. An important implication of this finding is that online sexual risk prevention and mental health are intertwined public health issues that must be considered in tandem. While the association between sexual risk-seeking behavior and mental health problems in offline contexts is well-established (3; 122), less work (127) has acknowledged that online sexual risk-seeking behavior may be an actionable way that some teens have found to cope given their mental health concerns. Therefore, instead of focusing on restricting teens' online sexual behavior and/or punishing them for it, we should expand youth mental health services so that youth have the support they need to make healthy sexual decisions both online and offline.

On the other hand, our findings also uncovered some legitimate scenarios where adolescents engaged in consensual sexting with other teens that was not motivated by mental health concerns or necessarily harmful to teens. For instance, LGBTQ+ youth use online peer support platforms to make one-to-one connections for support; yet, their sexual identities are often confounded with the types of support they seek in online spaces (174). Therefore, as we saw, support and intimacy may coincide, making such interactions appear more "risky" to outsiders who do not understand the challenges youth face when attempting to understand and explore their sexual identities. Similarly, our analysis uncovered how teens are expanding their dating horizons to foster new relationships online that they might not have had the opportunity to form in their local circles. Researchers argue that online relationships may benefit youth (288), particularly those who have difficulty forming romantic relationships, such as those who are on the autism spectrum (218). Therefore, we recommend future research take a broader, more nuanced perspective that includes both positive and negative aspects of online sexual experiences for adolescents, rather than only considering risks. In particular, our work highlights how sexting may be considered a normative behavior within healthy adolescent relationships but can quickly become problematic when one party believes the interaction is consensual, while the other party feels pressured into it.

As such, our findings suggest a need to move beyond assessing risks and negative outcomes of teens' sexual behaviors and work toward a better understanding of the benefits of sexting. In the case of consensual sexting with strangers, online anonymity and accessibility (332) might make it easier for teens to explore their sexuality and connect with other teens with the same sexual orientation. As coming out to families is usually difficult for adolescents (108), online platforms provide a space for teens to discuss these issues with peers. On the other hand, consensual sexting with strangers carries potential risks. For example, posts in our dataset revealed how online anonymity allowed teen users to seek out older strangers for sexting, in some cases by presenting themselves as older than their actual age—either to deceive other users or to evade platform age restrictions. This suggests measures taken by social media platforms to limit access by minors may be insufficient if they rely solely on self-reported age. In general, it would be helpful to transition the conversation away from unequivocally viewing all adolescent sexting behavior as negative to identifying the pathways in which it can be done safely and beneficially. The multifaceted nature of online sexual risk experiences should thus be considered in the research and design of social media platforms to better serve adolescent users.

3.5.4 Implications for Design

Based on our results, we make the following design-based recommendations for online sexual risk detection, prevention, and mitigation for teens:

3.5.4.1 Algorithmic Sexual Risk Detection Systems

We recommend that, if automated risk detection systems are deployed in real-world contexts, they take into account contextual features, such as levels of consent and relationship types when determining how best to support adolescents in navigating online sexual risk experiences. For instance, any indication that an online sexual exchange is non-consensual, regardless of the type of relationship, could immediately nudge a teen to take protective measures and/or seek help. Even consensual sexual experiences between teens and a stranger and/or family member may also indicate that risk mitigation procedures are needed. In the case that a sexual exchange is consensual, even among friends and/or dating partners, but the context indicates mental health problems (e.g, depression, suicide, self-harm), different risk mitigation strategies (e.g., trauma-informed) may be more appropriate. Further, if consensual sexting results in non-consensual sharing of sexual imagery, measures could be taken to prevent unauthorized distribution to third-parties. For instance, Meta recently announced that their private messaging platforms will be implementing end-to-end encryption and notify users when recipients of their disappearing messages take screenshots (Zuckerberg). While such privacy features may help discourage unauthorized sharing of sensitive content, researchers also need to carefully evaluate whether such features may unintentionally harm vulnerable users who take screenshots to document their abuse. Finally, if any sexual abuse of a minor is detected online, platforms should have specific procedures in place to proactively report such situations to the proper authorities for immediate investigation. Recent legislation makes online platforms culpable for sex trafficking that occurs on their platforms (230); similarly these platforms should also bear responsibility for sexual abuse propagated on their sites.

3.5.4.2 Intelligent Defaults for Sexting with Strangers

Based on our findings, we recommend future research investigate the feasibility and user acceptance of an "opt-out" privacy default, where social media platforms block strangers from privately contacting minors; however, these platforms should also give teens the ability to override this default. Further, risk awareness notifications should be included to inform teens about the potential benefits and risks associated with sexting with strangers. This design would be more powerful in providing both safety notices and choices (298) for teens to proactively manage their private interactions with strangers, rather than blocking strangers after receiving unsolicited sext messages.

3.5.4.3 Age Verification Systems

Another topic that emerged from our analysis was that of age, both in terms of teens engaging in sexually risky behavior with adults online and/or posing as adults to engage in sexual experiences with unknowing others. While researchers in the European Union (EU) have recently advocated for age verification systems (of Economics and Science) to ensure that internet users are of age to legally consent to the terms of use for various online platforms, we extend the recommendation of age verification as a way to ensure teens are not engaging with or as adults in online sexual exchanges that could potentially lead to illegal sexual activities involving minors. Such systems may help protect teens from "catfishing" attempts from adults posing as teens and protect adults from mistakenly getting sexually involved with a minor. Further research could focus on building efficient age verification systems and studying the feasibility of integrating these systems into online platforms to help implement ethical and personalized restrictions.

3.5.4.4 Designing for Computer-Mediated Consent

We also encourage more scholarship examining and designing better models for facilitating computer-mediated consent (390) by design, rather than leaving room for ambiguity. For instance, dialogues to determine whether a sexual exchange is consensual could include lightweight mental health evaluations to raise a teens' self-awareness of their underlying motivations to sext. Such self-reflections have been found to be a helpful approach for helping teens make better-informed decisions in other risk contexts. (94). Consent-based dialogues could also be embedded in the process of sharing of explicit content. General Data Protection

Regulation (GDPR) (291) requires online platforms that operate in the EU to give users the right to remove their personal data, but this does little to enable teens who share explicit imagery with others to retroactively withdraw consent. One potential area for future research would be to explore the use of blockchain technologies (115), so that teens (and other internet users) can protect, manage, and remove their sexually explicit digital trace data from the internet. By creating intentional models for handling the complexities of computer-mediated consent, researchers and practitioners can take the needed strides to prevent online child sexual exploitation.

3.5.5 Limitations and Future Work

We recognize some limitations in this work. The posts used in this study were written by adolescents on a single, albeit large, mental health peer support platform. Therefore, the generalizability of this study may be limited given the nature of the platform. We likely came across more negatively sexual experiences and abuse disclosures based on the platform's purpose and norms, which may not be as prevalent on general purpose social media platforms. Future research should take into account different platforms to verify that our results can be replicated to a more diverse population of adolescents. Secondly, while we were able to build and train machine learning classifiers and deeply analyze self-disclosures of online sexual risk experiences, we did not have the victims' own interpretations of these experiences, i.e., how risky they perceived these experiences to be an aspect that can be valuable in training machine learning-based risk detection models (180). Therefore, future research should consider adolescents' perspectives on their disclosures of online sexual risk experiences by asking them to flag or comment on their own posts. Third, since we trained the models after combining friends and dating partners as one class, the models classified it as one class for the rest of the dataset. Therefore, we might have lost some qualitative differences between friends and dating partners that future research can investigate within a large dataset.

Conducting studies of sensitive data from a vulnerable population is a critical matter within the social computing and HCI communities (67; 117). While the platform's terms of service stated that the posts may be used for research purposes, we took more precautionary measures to protect youth from potential harm. We made sure to anonymize the data and paraphrase example posts to prevent them from being publicly searchable or traceable to a specific individual. Privacy and ethics in the context of this type of research, including the suggested design implications above, need to be persistent topics of discussion in the years coming.

CHAPTER 4

STUDY 2: Profiling the Offline and Online Risk Experiences of Youth to Develop Targeted Interventions for Online Safety

Citation: Ashwaq Alsoubai, Afsaneh Razi. Zainab Agha, Shiza Ali, Gianluca Stringhini, Munmun De Choudhury, Pamel J Wisniewski. Profiling the Offline and Online Risk Experiences of Youth to Develop Targeted Interventions for Online Safety. Proceedings of the ACM On Computer-Supported Cooperative Work And Social Computing (CSCW '24).

We conducted a study with 173 adolescents (ages 13-21), who self-reported their offline and online risk experiences and uploaded their Instagram data to our study website to flag private conversations as unsafe. Risk profiles were first created based on the survey data and then compared with the risk-flagged social media data. Five risk profiles emerged: Low Risks (51% of the participants), Medium Risks (29%), Increased Sexting (8%), Increased Self-Harm (8%), and High Risks Perpetration (4%). Overall, the profiles correlated well with the social media data with the highest level of the risk occurring in the three smallest profiles. Youth who experienced increased sexting and self-harm frequently reported engaging in unsafe sexual conversations. Meanwhile, high risks perpetration was characterized by increased violence, threats, and sales/promotion of illegal activities. A key insight from our study was that offline risk behavior sometimes manifested differently in online contexts (i.e., offline self-harm as risky online sexual interactions). Our findings highlight the need for targeted risk prevention strategies for youth online safety.

4.1 Introduction

Modern-day youth are found to experience a myriad of online and offline risks such as substance misuse (2), cyberbullying (27), or unwanted online sexual experiences, including exposure to sexually explicit materials (219). The research community has shown a great interest in understanding the boundaries between the offline and online behaviors of youth, mainly in the context of these risk experiences (106). Offline and online risk experiences of youth were found to be driven by the same underlying factors related to the general propensity to experience risks (135). There is well-established research that has focused on examining the correlation between youth online and offline risk experiences such as between online sexual risks and self-harm (242; 353) or substance misuse (40), or between online harassment and self-harm (311) or substance misuse (275; 380). This line of research has examined such correlations by primarily conducting survey-based and longitudinal studies that used self-reported data from youth, which could be subject to recall and social desirability biases (334). To overcome these biases, Pinter et al. (276) called for using social media trace data to have an in-depth understanding of youths' lived experiences online. Additionally, while the current efforts for online safety education and prevention initiatives (258; 119) are valuable in raising youth awareness about online safety, such efforts are often designed for general populations of youth, rather than being tailored to the unique differences and needs of subgroups of youth who share similar offline and online experiences. Therefore, there is a need to personalize online safety prevention and education strategies for youth to improve these outcomes. Profiling youth based on their similar risk experiences is one way to achieve this level of personalization. In this paper, we addressed the following research questions:

RQ1: *What are the differing profiles of youth based on their self-reported offline and online risk behaviors (i.e., online harassment, sexting, self-harm, and offline risk behavior)?*

RQ2: *How do these risk profiles correlate and/or differ when compared to youths' risk-flagged private Instagram trace data?*

RQ3: *How do unsafe Instagram conversations differ linguistically between the different risk profiles?*

To answer RQ1, we conducted a Mixture Factor Analysis (MFA) to create youth profiles based on self-reported offline and online risk experiences. Profiling is one of the common approaches to identify groups of human or nonhuman subjects based on analyzing the correlations between data (156). For RQ2, we conducted a between-group analysis using a (χ^2) test to uncover key differences between the youth profiles based on their self-risk-flagging behaviors. We then used machine learning generative models of text to identify the unique linguistic differences between the profiles' unsafe conversations to answer RQ3. Through these analyses, we identified five unique youth risk profiles: 1) Low Risks (51% of the participants), 2) Medium Risks (29%), 3) Increased Sexting (8%), 4) Increased Self-Harm (8%), and 5) High Risks Perpetration (4%).

In summary, this work makes the following novel research contributions:

- It identifies correlations between youths' self-reported responses and their Instagram digital trace data to address the methodological gaps within the literature, which contributes to the body of works related to youth online safety.
- It illustrates the importance of acknowledging the multi-dimensional nature of youth online and offline risk behaviors to identify unique clusters of youth.
- It improves our understanding of youth risk assessments and perceptions by aligning their explicit risk flagging to their self-reports, the evidence of which can help provide youth more agency for their online/offline safety.

- It demonstrates how some offline behaviors of youth were manifested differently online. We provide implications for designing safer online interactions for youth based on their different profiles.

Our work aims toward building a comprehensive understanding of how youth experience offline and online risks by triangulating their self-report survey data with their private social media data.

4.2 Background

In this section, we synthesize literature on youth safety and highlight potential gaps that motivate this work to delineate profiles of youth and contextualize their self-reports of risks with social media trace data.

4.2.1 Relations between Offline and Online Risk Behaviors of Youth

A major line of research has primarily focused on how online behaviors manifested offline (353; 36; 380; 295). For instance, Wachs et al. (353) conducted a survey with 2506 adolescents (ages 13–16 years) and found that online sexual risks (i.e., pressured sexting) were positively and significantly correlated with offline non-suicidal self-harm. Yoon et al. (380) have also surveyed 10th-grade students (N = 2,768) and after the 12-month follow-up, they found that youth who experienced cyberbullying, regardless of their roles as witnesses, perpetrators, or victims, had higher odds of substance misuse. This line of research has mainly not considered the offline and online risk behaviors in reverse or considered a bi-directional influence on each other. In addition, most of these studies have taken a unidimensional approach by studying certain offline and online risk behaviors in the absence of others, which is a research gap that will be discussed in detail in the following section.

Identifying the positive correlations between the myriad of offline and online risk behaviors of youth means that the scholars in this line of research aggregate responses of youth risk experiences as a single scale. By summing these responses into an overall measurement of risk experiences, these scholars have made an assumption of the unidimensionality of these risks (i.e., they assume that all youth experience the same risks or level of risks) (126). This approach allows them to inform about the covariates of offline and online risks. In doing so, they might oversimplify the actual risk experiences of youth. In fact, findings from several youth sample studies point to differences in risks that youth encounter, particularly regarding the risk prevalence and types (96; 121; 31). As such, in this paper, we extend this line of research and address this gap by illustrating how offline and online risk experiences of youth are in fact multidimensional, i.e. that different youth could experience different types of risks across offline and online contexts. This illustration will be presented in this study by creating profiles of youth based on their self-reported offline and online risk experiences. To our knowledge, this study is the first or one of the first studies that takes an additional step forward to present an empirical approach for better understanding the multidimensional nature of youth online and offline risks as well as how well youths' self-reported risks are aligned with their social media data.

4.2.2 Profiling Youth to Understand their Offline and Online Risk Experiences

There is a growing body of knowledge within youth safety literature that has attempted to elucidate unique profiles of adolescents to better understand the heterogeneity in this population based on their distinct risks experiences (325; 152; 45; 182). For instance, related studies have presented profiles of youth based on risks that were encountered exclusively either online or offline. Using Latent Class Analysis (LCA), Bishop et. al (45) identified four profiles of substance misuse for gang-involved youth: Non-Users (38%), Past Users (15%), Casual Users (27%), and Frequent Multi-Users (21%). These profiles revealed a nuanced understanding of the differences among the gang-involved youth in their substance misuse along with the ecologies that either promoted or inhibited certain patterns of misuse, against the common perceptions that all youth in this population are users. Recently, emergent works have acknowledged the importance of incorporating the offline and online context of risks when creating youth profiles to provide a more holistic understanding of such risks. For instance, Kim et al. (177) created profiles for adolescents based on their offline and online bullying behaviors. Four profiles emerged: (1) Low Risk (85.3%), who reported the lowest levels of engaging in both offline and cyberbullying, (2) High Risk (2.4%), who showed high levels of engagement in both bullying and victimization online and offline, (3) Offline Risks (5.1%), who had high scores for offline bullying, but low scores for cyberbullying, and (4) Online-Risk Group (7.2%), who reported high scores of engagement in the online domain, but low scores in offline bullying. Through these profiles, they were able to confirm the co-occurrence in the roles of bullying (i.e., victim and perpetration) across the online and offline contexts. These studies highlight the value of examining youths' heterogeneity related to their risk experiences, which would help scholars and practitioners to better understand the dynamics of risks in the youth population and therefore delineate targeted and evidence-based intervention initiatives and education plans. Our work builds upon this literature by adopting a similar approach to profile youth based on their distinct youth risk experiences. We contribute to the literature by moving beyond a narrow view of a subset of related risks to studying a wider array of risks, including offline risk behaviors, offline self-harm, unwanted online sexual risks, online harassment roles (e.g., perpetrator vs. victim), and online sexting to holistically investigate the diversity of risks in youth profiles.

4.3 Methods

Below, we give an overview of our study, followed by a detailed account of our research methods.

4.3.1 Study Overview

We developed a secure, web-based system, where participants first completed a web-based survey; then, they were asked to login into their Instagram accounts to download their Instagram data files. We selected Instagram because it is the most popular social media platform after YouTube and TikTok among youth (27). As Pew Research recently found that six in ten teens engaged as active users of the platform (27). Instagram also enables users to easily download their data based on the General Data Protection Regulation (GDPR) (128), which mandates social media companies to allow users to download their own personal data. After uploading their Instagram data to our secured system, participants were asked to review their private message conversations, flag messages that made them or someone else feel uncomfortable or unsafe, and provide contextual information (e.g., what happened, with whom) about the interaction. In the subsections below, we provide more details about the survey constructs and Instagram data donation procedure.

4.3.2 Survey Design

To measure youth risk experiences, we utilized pre-validated survey measures, including Risky Behavior Questionnaire (28), Inventory of Statements About Self-harm (183), Cyber-Aggression Victimization (315), Cyber-Aggression Perpetration (315), Unwanted Sexual Solicitations and Approaches (235), and Youth Produced Sexual Images (Sexting) (235) (Appendix A). The Likert-scale of these measures was from 1-5 (1-Never, 2- Rarely, 3-Sometimes, 4- Often, 5- All the time). We slightly updated the wording of the question for the online risk experience constructs to ask specifically about the participants' experiences when using Instagram, rather than in general. Below, we explain the measures in more detail.

4.3.2.0.1 Offline Risk Behaviors.

In this study, we leveraged the Risky Behavior Questionnaire (RBQ) scale by Auerbach and Gardiner (28) to measure the frequency of a myriad of offline risk behaviors during adolescence, including substance misuse, unsafe sex, cheating, and gambling. Using this construct gave us a more holistic view of youth risk experiences rather than only focusing on their online behaviors. Non-suicidal self-harm behavior has become more prevalent recently, especially among vulnerable populations like youth (221); therefore, we included the Inventory of Statements About Self-harm (ISAS) scale by Klonsky and Glenn (183), which was designed to understand the non-suicidal self-harm behaviors. This scale quantifies the frequency of intentional youths' self-harming behaviors, including cutting, scratching, and hitting away from suicidal reasons.

4.3.2.0.2 Online Risks behaviors.

We included The Cyber-Aggression Victimization (CAV) scale developed by Shapka and Maghsoudi (315), which measured youth experiences of online harassment in different forms such as receiving hurtful comments, gossip about them, or having an embarrassing post, photo, or video on Instagram. To better understand youth risk experiences, we not only used the online harassment construct of youth as victims, but also included the online harassment perpetration, where the youth were the perpetrators. To do this, we adopted the Shapka and Maghsoudi (315) scale for Cyber-Aggression Perpetration (CAP) to measure the prevalence of online harassment experiences that youth committed online. The questions were similar to the Cyber-Aggression Victimization scale, but were rephrased to be about the participants committing harassment instead of being victims. We utilized the Unwanted Sexual Solicitations and Approaches scale from the Youth Internet Safety Survey (YISS) developed by Mitchel et al. (235) to measure the unwanted online sexual risks, including sexual messages, requests to engage in sexual activities and/or sexual conversations, and unexpected exposures to nude pictures or people having sex. This scale combines two scales, which were the Unwanted sexual solicitations and approaches and Unwanted exposure to pornography, into one scale called Unwanted Online Sexual Risks. We did not include the harassment questions to avoid repeating the cyberbullying questions from the CAV and CAP scales, which were more comprehensive of the online harassment experiences. We also used Mitchel et al.'s (235)Youth Produced Sexual Images (Sexting) construct from the same YISS. This scale consists of five questions three of them about the possession and distribution of digital imagery depicting nudity of a minor (under the age of 18). Since the possession and distribution of such materials is considered a federal crime, we did not ask our participants these questions in the survey. Therefore, this measure will mainly quantify the production and distribution of youth sexual imagery or videos, particularly whether they send or receive any personal nude/semi-nude media (pictures or videos).

Type	Measures	No. Items	Cronbach's alpha	Mean	SD	Skewness	Kurtosis
Risk Experiences	**Offline Risk Behaviors**	20	0.82	1.54	0.38	1.2	2.71
	Offline Self-harm	12	0.85	1.55	0.62	1.58	2.14
	Unwanted Online Harassment	12	0.9	1.99	0.72	1.01	0.43
	Online Harassment Perpetration	12	0.9	1.28	0.44	2.62	7.21
	Unwanted Sexual Experiences	5	0.84	2.18	0.88	0.36	-0.69
	Interpersonal Online Sexting	2	0.76	1.4	0.73	2.24	0.43

Table 4.1: Measures descriptive statistics.

Table 4.1 summarizes the descriptive statistics of the pre-validated constructs used in this study. Cronbach's α's, which measures the internal consistency of survey constructs (89), was higher than the acceptable 0.7 threshold (79). More importantly, the Interpersonal Online and the Unwanted Online Sexual scales remained reliable after the changes we did with Cronbach's α of 0.76 and 0.84 respectively. Table 4.2 demonstrates the significant positive correlations between the risk experience measures, which suggest that youth who encountered one type of risk experience were also more likely to encounter another.

Type	Measure	RBQ	ISAS	CAV	CAP	YISS	SEXT
Risk Experiences	**Offline Risk Behavior (RBQ)**	1					
	Offline Self-Harm (ISAS)	.466**	1				
	Unwanted Online Harassment (CAV)	.378**	.224*	1			
	Online Harassment Perpetration (CAP)	.283**	.183*	.367**	1		
	Unwanted Sexual Experiences (YISS)	.476**	.218*	.554**	.275**	1	
	Interpersonal Online Sexting (SEXT)	.399**	.204*	.393**	.234**	.506**	1

Table 4.2: The correlation between the risk experiences measures (Unwanted Online Sexual Risks, Interpersonal Online Sexting, Unwanted Online Harassment, Online Harassment Perpetration, Offline Self-harm, and Offline Risk Behaviors). All risk experiences measures were significantly positively correlated with each other. * p-value ¡ 0.05, ** ¡ 0.01, and *** ¡ 0.001

Lastly, the participants also answered demographic questions about their sex, age, race, and sexual orientation.

4.3.3 Instagram Data Collection

After the participants uploaded their Instagram files in the form of zipped JSON files to our system, their direct conversations were displayed back to them in reverse chronological order to review their past conversations. For the participants who are 18 years and older, their conversations when they were under 18 were shown to them first. The participants were asked to review and evaluate their past conversations as to whether they made them feel uncomfortable or unsafe. For each unsafe conversation, the participants were then asked to flag the unsafe messages and evaluate them based on the risk severity levels (i.e, low, medium, high), which were adopted based on prior literature (363), and risk types. The risk types included harassment, sexual messages or solicitations/nudity, hate speech/threat of violence, sale or promotion of illegal activities, digital self-injury, or spam, which were derived from Instagram reporting feature risk categories[1]. For the unsafe conversations, the participants were also asked to specify their relationship with the other person involved as a stranger, acquaintance, friend, family, or significant other. This will be mainly used to supplement the qualitative reading that will be presented in section (4.3) to contextualize the conversations better. Please be aware that throughout the study, we used the term "risky" to refer to uncomfortable or unsafe conversations.

4.3.4 Participant Recruitment and Demographics

For this study, participants were recruited based on the following selection criteria: 1) between 13 and 21 years old, 2) English speakers, 3) located in the United States, 4) had an active Instagram account for at least 3 months during the time they were a teen (ages 13-17), had direct conversations with at least 15 people, and 5) had at least two conversations made them feel unsafe or uncomfortable. Per the requirements of our Institutional Review Board (IRB), participants who were under the age of 18 were required to provide

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

parents' consent and their own assent prior to their enrollment to the study. Participants who were over 18 years old completed the adult consent form. During the study, we disclosed our status as mandated child abuse reporters and warned participants that any instances of child pornography would have to be reported to the proper authorities. Therefore, we clearly requested the participants to not upload any content that includes the nudity of minors and described to them the required steps to delete such content from their Instagram files prior to uploading to our study system. Additionally, we obtained a Certificate of Confidentiality from the National Institute of Health, which protects the participants' privacy and prevents the subpoenaing of the data during legal discovery. All personally identifiable information from any textual or image data was removed and all quotations were paraphrased in our results to further protect the youths' privacy. The participants were compensated with a \$50 Amazon gift card for their data and time after verifying the quality of their data.

For this study, ($N = 173$) youth participants were able to successfully complete both parts of the study. Most of the participants were females (67%), (23%) males, and (10%) non-binary. Half of the participants identified themselves as heterosexual or straight 50%, while the rest were bisexual (28%), homosexual (9%), and 13% preferred to self-identify. In order to examine the impact of these demographics on the generated youth profiles, we conducted between-group analysis (χ^2) between the profiles based on sex, age, and sexual orientation. From the 173 participants, we collected ($N = 33,469$) conversations and out of these conversations ($N = 32,256$) were flagged as safe conversations and ($N = 1,213$) as unsafe conversations. Out of these unsafe conversations, ($N = 3,066$) messages were flagged for risk levels and types. The following section presents the results of this study.

4.3.5 Data Analysis Approach

We combined the analysis of self-reported risk experiences and social media trace data from participants. We applied the Mixture Factor Analysis (MFA) to create the youth profiles based on online and offline risk experiences. Next, between-group analysis (χ^2) was performed to examine any significant differences between the profiles based on their risk flagging. Then, an unsupervised language modeling approach was performed to extract key linguistic differences in the profiles' unsafe private conversations. The following sections describe these approaches in more detail.

4.3.5.1 Mixture Factor Analysis (MFA) to Profile Participants' Self-Reported Risk Experiences (RQ1)

To create the youth risk profiles, we used the self-reported measures of youths' risk experiences that were explained in Section 3.2. We conducted a series of Mixture Factor Analyses (MFAs) with a robust maximum likelihood estimator to group like-minded youth. Mixture Factor Analysis creates profiles based on a "mixture" of factor mean scores of the self-reported measures (244). This approach is useful because it demonstrates the relationship between each factor (risk experience) for different groups of youth. Studying youth risk experiences based on the factors improves the interpretability and generalizability of the findings as we study key risk experiences patterns (factors) rather than many discrete risk behaviors (items) (184).

The MFA provides only indicators for the optimal number of profiles rather than explicit information to compare the relative quality of the resulting solutions with differing numbers of profiles. An optimal profile solution might exist with a maximum value for the Shannon entropy (314), a minimum value of Bayesian Information Criterion (BIC), which assesses the profile solution parsimony (214), or when the log likelihood starts to level off, especially that the higher number of profiles may increase the overall model fit, which might not be significant (184). These indicators may not agree on the optimal profile solution, which usually leads to also use the substantive grounds beside the fit measure indicators (255): the optimal cluster solution could be decided based on the interpretability and reasonability of the cluster distributions (e.g., avoid solutions with very small clusters and/or have very large cluster means). For this study, we thus leveraged both this substantive ground and the fit measures (in our case a maximum level of entropy and the log-likelihood levels off) were used to decide on the optimal number of the youth risk profiles.

Classes	BIC	Entropy	LL
	Risk Profiles		
2	1024.35	0.81	-473.68
3	915.43	0.85	-435.715
4	959.59	0.88	-412.424
5	**929.04**	**0.90**	**-382.716**
6	945.15	0.90	-376.33

Table 4.3: Risk MFA model fit statistics. The bold values indicate the optimal solution (5 profiles) based on the maximum levels of entropy and log likelihood.

To this end, table 4.3 lists the resulted profile solutions of the MFA based on different numbers of profiles. For the youth risk profiles, we did not observe any substantial improvements beyond a 5-cluster solution. The BIC, reached a minimum value for the 3 and 5 cluster solution. For the 5-cluster solution, the entropy levels reached its maximum value and the log likelihood started to taper off after a 5-cluster solution. Thus, for this study, a 5-cluster solution was the optimal solution for youth risk profiles. To confirm the existence of significant differences between the profiles based on our participants' risk experiences, we conducted a series of ANOVA tests (90) with the risk experiences: offline risk behaviors, offline self-harm, unwanted online harassment, online harassment perpetration, unwanted sexual experiences, interpersonal sexting, as dependent variables and the generated profiles as the independent variable. We also applied a series of post-hoc tests, to compare individual profiles with one another (226). The identified differences provide a comprehensive overview of the distinct risk experiences of the youth profiles.

4.3.5.2 Between Group Analysis based on the Risk Flagging of Youth Profiles (RQ2)

To correlate profiles' self-reported risk experiences and actual risk flagging behaviors and answer RQ2, between-group analysis (χ^2) was performed between the youth risk profiles based on the risk levels and risk types of the participants' flagged messages ($N = 3,066$). χ^2 tests of independence are between-group tests which are used for two or more variables with normal distribution (316). The standardized residuals, which are calculated by "dividing the product of subtracting expected from observed values by the square root of the expected value" (316), were used in this study to show the significant associations between the youth risk profiles and their risk flagging. Through these χ^2 tests, we mapped youth risk profiles self-reported responses to their actual risk flagging (risk levels and types) of their unsafe private messages.

4.3.5.3 Linguistic Differences in Youth Risk Profiles' Social Media Conversations (RQ3)

To answer RQ3, we used an unsupervised language modeling technique, Sparse Additive Generative Model (SAGE) (111) to examine the linguistic differences in the youth risk profiles' unsafe conversations. SAGE extracts distinguishing keywords in given texts by comparing the parameters of logistically parameterized multinomial models using self-tuned regularization to control the trade-off between frequent and rare terms (111). We applied SAGE to identify *n*-grams (*n*=1,2,3) that differentiate the unsafe Instagram private conversations flagged by the five youth risk profiles (after merging the conversations for each profile). The SAGE value of an *n*-gram indicates the level of its "uniqueness." SAGE results were packed up with content analysis (158) to contextualize the extracted keywords and better understand its context within the profiles' unsafe conversations. Using SAGE and the qualitative reading enabled us to capture the distinctive and salient characteristics of the profiles' unsafe conversations content to compare them, contextualize their risk flags, and holistically understand their self-reported responses to the online/offline risk experiences survey.

4.4 Results

In this section, we presented the created profiles based on the self-reported risk experiences (RQ1), the types and levels of risks these profiles flagged (RQ2), and the linguistic differences in their unsafe private Instagram conversations (RQ3).

4.4.1 Youth Self-Reported Risk Profiles (RQ1)

The resulting five MFA clusters represented youth risk profiles that described a set of distinct risk experiences that the members of each cluster encountered online and offline. Figure 4.1 shows the profiles along with the percentages break down of number of participants who were members of each profile. The profiles' average scores of the risk experience constructs were mainly ranged from 1 to 3.5, where the highest reported scores were used to semantically label the profiles. In the web graph, the constructs are shown clockwise in descending order by frequency from experienced the most to the least by the aggregated average scores of all of the groups. Table 4.4 shows that ANOVA yielded significant differences between the youth profiles based on the risk experience constructs. The following section describes these profiles along with the significant differences in detail.

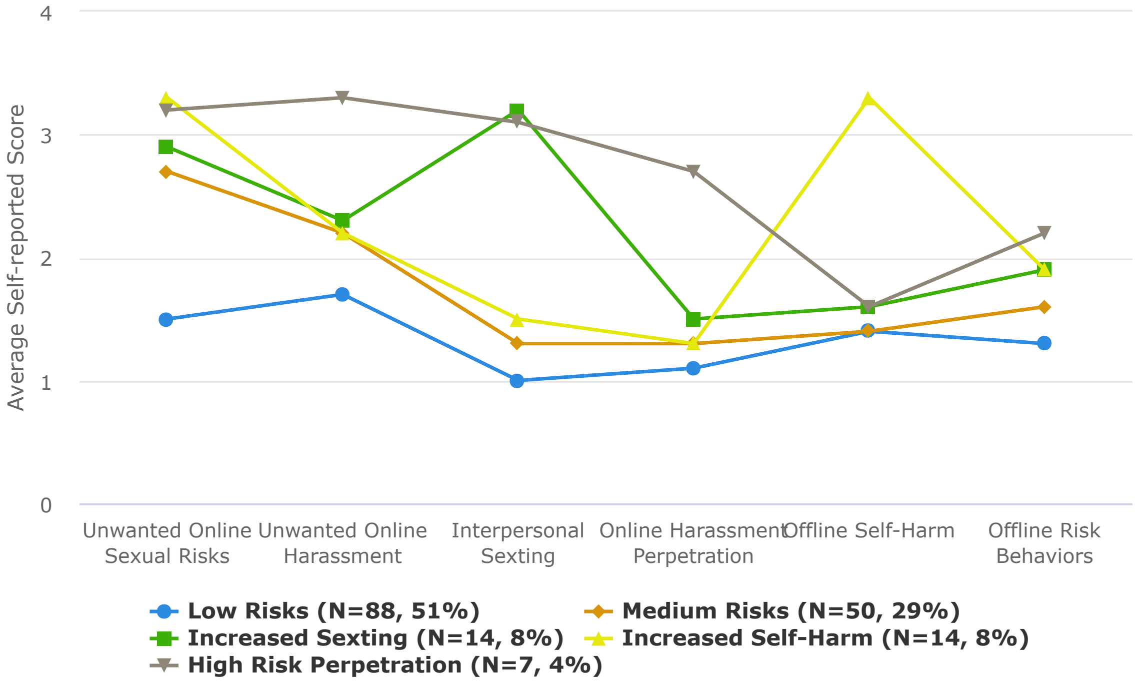


Figure 4.1: Youth risk profiles (N=173).

4.4.1.0.1 Low Risks (51%):

This profile represented the largest group forming (51%) of our participants. Based on the ANOVA results listed in Table 4.4, participants in the Low Risks profile self-reported significantly lower scores for all the risk experience measures online and offline, comparing with other profiles. Overall, Low Risks profile encountered less risk experiences online and offline as this profile's responses mostly fell between "Rarely" and "Never" on the Likert scale.

4.4.1.0.2 Medium Risks (29%):

As the second largest group, youth in the Medium Risks profile reported middle level scores for the online and offline risk experience measures, compared to other profiles. Looking at the ANOVA results Table 4.4, we found that Medium Risks profile reported significantly higher average scores than the Low Risks profile

Constructs	df	F	*p*-value	Significant Pairwise Differences (Mean)
Unwanted Online Sexual Risks	4	92.11	$p < 0.001$	Medium Risks (m=2.77), Increased Sexting (m=2.94), Increased Self-harm (m=3.35), and High Risks Perpetration (m=3.21) >Low Risks (m=1.52) Increased Self-harm (m=3.35) >Medium Risks (m=2.77)
Unwanted Online Harassment	4	27.58	$p < 0.001$	Medium Risks (m=2.25), Increased Sexting (m=2.33), Increased Self-harm (m=2.91), and High Risks Perpetration (m=3.21) >Low Risks (1.60)
Interpersonal Sexting	4	128.80	$p < 0.001$	Increased Sexting (m=3.22), Medium Risks (m=1.35), Increased Self-harm (m=1.59), and High Risks Perpetration (m=3.1) >Low Risks (m=1.06) Increased Sexting (m=3.22) >Medium Risks (m=1.35)
Online Harassment Perpetration	4	50.11	$p < 0.001$	High Risks Perpetration (m=2.69) >Low Risks (m=1.17), Medium Risks (m=1.30), Increased Sexting (m=1.54), and Increased Self-Harm (m=1.39)
Offline Self-harm	4	48.64	$p < 0.001$	Increased Self-Harm profile (m=3.32) >Low Risks (m=1.43), Medium Risks (m=1.48), Increased Sexting (m=1.65), and High Risks Perpetration (m=1.67)
Offline Risk Behaviors	4	16.38	$p < 0.001$	Medium Risks (m=1.68), Increased Sexting (m=1.90), Increased Self-Harm (m=1.99), and High Risks Perpetration (m=2.23) >Low Risks (m=1.38)

Table 4.4: ANOVA results and the summary of significant pairwise differences. There were significant differences between the youth profiles based on the listed constructs (*p*-values less than 0.05).

for the unwanted online sexual risks, unwanted online harassment, interpersonal sexting, and offline risk behaviors. We also found that Medium Risks profile experienced significantly less unwanted online sexual risks than increased self-harm and less interpersonal sexting than the Increased Sexting profile. On the scale, this profile's responses mostly fell between "Sometimes" and "Rarely" for all the measures.

4.4.1.0.3 Increased Sexting (8%):

Compared to other profiles, youth in the Increased Sexting profile self-reported the highest levels of interpersonal sexting (mostly "Sometimes" or "Often" on the scale). An ANOVA (Table 4.4) yielded that the youth in this profile experienced significantly more frequent interpersonal sexting than Low and Medium Risks profile. It was clear that this profile also experienced incidents of unwanted sexual risks and unwanted online harassment, which were reported on the scale as "Sometimes" or "Rarely." This profile rarely experienced online risk perpetration, offline self-harm, and offline risk behaviors.

4.4.1.0.4 Increased Self-Harm (8%):

An ANOVA (Table 4.4) showed that the Increased Self-Harm profile self-reported significantly more frequent offline self-Harm experiences than all profiles. An ANOVA also revealed that this profile had significantly more unwanted online sexual risks than Low and Medium Risks profiles. Youth in this profile reported between "Often" and "Sometimes" on the scale for the offline self-harm and "Sometimes" for the unwanted online sexual risks.

4.4.1.0.5 High Risks Perpetration (4%):

High Risks Perpetration represented the smallest profile. In comparison to other profiles, high risks perpetration profile reported the highest average scores for online harassment perpetration as their responses fall mostly in "Sometimes" on the scale. Looking at the ANOVA results (Table 4.4), the High Risks Perpetration profile had significantly higher online harassment perpetration experiences than all profiles. Overall, this profile experienced significantly more online and offline risk experiences than Low Risks except that this profile reported significantly less levels of offline self-harm experiences than the Increased Self-Harm profile.

Demographics	**Parameter**	**Low Risks**	**Medium Risks**	**Increased Sexting**	**Increased Self-Harm**	**High Risk Perpetration**	χ^2
Sex	Female	32%	23%	5%	5%	2%	p-value = 0.24
	Male	14%	5%	2%	1%	2%	
	Non-Binary	4%	2%	1%	2%	0%	
	Prefer to self-identify	2%	0%	0%	0%	0%	
Age	13-15	15%	7%	1%	1%	1%	p-value = 0.12
	16-18	23%	13%	5%	6%	2%	
	19-21	14%	9%	2%	0%	0%	
Sexual Orientation	Heterosexual or straight	29%	14%	4%	1%	2%	p-value = 0.08
	LGBTQ+	22%	15%	5%	6%	2%	

Table 4.5: Distribution of demographics by the risk profiles and χ^2 results. % Out of the total number of participants (N=173).

Overall, the youth profiles highlight the multidimensionality of youth risk experiences. When analyzing these profiles based on the reported demographics, we found that the profiles were not impacted by the youths' demographics (there were no significant differences yielded between the youth profiles based on sex, age, and sexual orientation using χ^2 as listed in table 4.5). This result is noteworthy as it suggests that profiling youth based on their risk experiences, rather than their demographic characteristics, yields additional insight that may be missed if we focused on demographic information alone.

4.4.2 Youth Risk Profiles Significantly Differed Based on the Flagged Risk Levels and Types (RQ2)

Next, we examined whether the self-reported risk experiences had any relationship with the youths' risk-flagged social media data. The χ^2 test uncovered key differences between the youth profiles and their flagged risk messages based on risk severity levels and types, which will be presented in the following sections.

4.4.2.1 Youths' Flagged 'Risk Levels' Aligned with their Self-Reports.

The χ^2 test indicated a significant association between the youth profiles and their flagged risk levels ($\chi^2(df = 8, N = 3,066) = 94.38, p < 0.001$). As illustrated in Figure 4.2, a strong positive association was found between the Low Risk profile and the number of conversations flagged as low risk level in the social media data. Further, a significant negative association was found with medium and high risk levels. This indicated that youth in this profile were most likely to flag their unsafe messages with a low risk level, which clearly aligned with their low average scores for the self-reported risk experiences. For the Medium Risks profile, by looking at the standardized residuals in Figure 4.2, we found a significant positive association between the Medium Risks profile and medium risk level and a significant negative association with high risk level. This suggested that the unsafe messages of this profile were most likely flagged as medium risk. This also showed an alignment between their medium scores for the self-reported risk experiences and their risk level flagging.

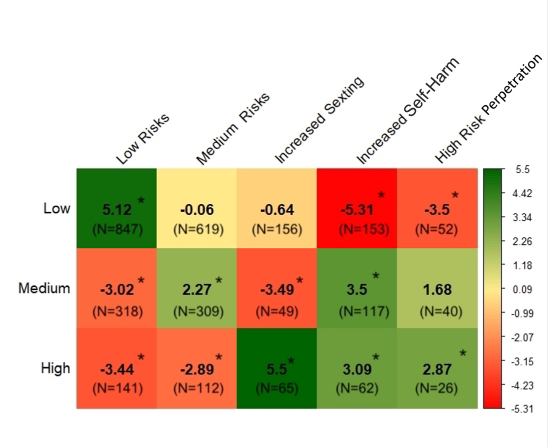
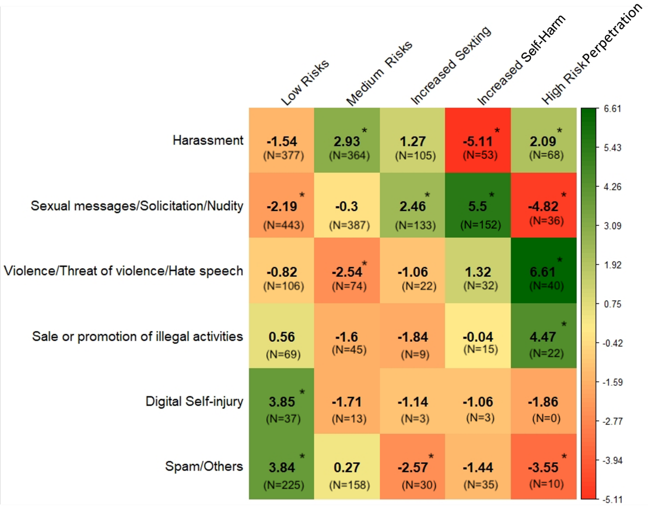


Figure 4.2: Results (standardized residuals) of the between group analysis for risk levels of the risk profiles ($N = 3,066$). (*) indicates significant association. Note that green denotes a positive association, while red denotes a negative one.

For the rest of the profiles (Increased Self-Harm, Increased Sexting, and High Risks Perpetration), the standardize residuals showed significant positive associations between these profiles and the high risk level as demonstrated in Figure 4.2. This finding suggested that the unsafe messages flagged by these profiles were more likely to be a high risk level. A significant negative associations were found between the Increased Sexting profile and medium risk level, suggesting that adolescents in this profile were less likely to flag their unsafe messages as medium level. On the contrary, the Increased Self-Harm profile showed a significant positive association with medium risk level, which indicated that the unsafe messages of this profile were more likely to be high and/or medium risk levels. A significant negative associations were found between the Increased Self-Harm and High Risks Perpetration profiles and low risk level, which suggested that these profiles were less likely to have unsafe messages with low risk level. These findings show a clear alignment between these profiles and their self-reports as each profiles reported the highest average scores for certain risk experiences (displayed at Figure 4.1).

4.4.2.2 Youth's Flagged 'Risk Types' Mostly Aligned with their Self-Reports.

The youth risk profiles were significantly different based on their flagged risk types using χ^2 test ($\chi^2(df = 20, N = 3,066) = 167.43, p < 0.001$). Looking into the standardized residuals in Figure 4.3, a strong positive association was found between Low Risks profile and risk types including digital self-injury and spam/others, along with a significant negative association with sexual messages/solicitation/nudity risk. This suggested that when adolescents in Low Risks profile flagged their messages, they most likely flagged them as digital self-injury or spam/others and less likely to flag for sexual messages/ solicitation/nudity. Generally, it was not surprising to see this profile mostly flagged spam/others, which matched with their self-reports and low risk level flag; however, finding the digital self-injury as part of their flagging warranted further qualitative unpacking, which will be done in section 4.3. For the Medium Risks profile, a significant positive association was found between this profile and harassment and a significant negative association with hate speech/threat of violence as shown in Figure 4.3. This finding suggested that the risk messages of this profile were more likely to be flagged as harassment and less likely to be flagged as hate speech/threat of violence. This group self-reported medium levels scores for the unwanted online harassment, which suggested a fair alignment between their self-reports and risk flagging.



	Low Risks	Medium Risks	Increased Sexting	Increased Self-Harm	High RiskPerpetration
Harassment	-1.54 (N=377)	2.93* (N=364)	1.27 (N=105)	-5.11* (N=53)	2.09* (N=68)
Sexual messages/Solicitation/Nudity	-2.19* (N=443)	-0.3 (N=387)	2.46* (N=133)	5.5* (N=152)	-4.82* (N=36)
Violence/Threat of violence/Hate speech	-0.82 (N=106)	-2.54* (N=74)	-1.06 (N=22)	1.32 (N=32)	6.61* (N=40)
Sale or promotion of illegal activities	0.56 (N=69)	-1.6 (N=45)	-1.84 (N=9)	-0.04 (N=15)	4.47* (N=22)
Digital Self-injury	3.85* (N=37)	-1.71 (N=13)	-1.14 (N=3)	-1.06 (N=3)	-1.86 (N=0)
Spam/Others	3.84* (N=225)	0.27 (N=158)	-2.57* (N=30)	-1.44 (N=35)	-3.55* (N=10)

Figure 4.3: Results (standardized residuals) of the between group analysis for risk types of the youth risk profiles ($N = 3,066$). (*) indicates significant association.Note that green denotes a positive association, while red denotes a negative one.

For the Increased Self-Harm and Sexting profiles, the χ^2 test yielded strong positive associations between these profiles and the sexual messages/solicitations/nudity risk type as shown in Figure 4.3. This result indicated that the risk messages flagged by these profiles were mostly flagged as sexual messages/solicitations/nudity. For the Increased Sexting profile, a significant negative association was found between this profile and spam/others risk type, which suggested that this profile was less likely to flag their messages for spam/others. While for the Increased Self-Harm profile, there was a negative significant association between this profile and harassment risk type, which suggested that this profile was less likely to flag their messages for harassment. These findings aligned with their self-reports since both of these profiles self-reported higher levels of unwanted online sexual risks. For the High Risks Perpetration profile, Figure 4.3 shows a significant positive association between this profile and risk types including harassment, hate speech/threat of violence, and sales or promotion of illegal activities and a significant negative association with sexual messages/solicitations/nudity and spam/others risk types. This indicated that youth in this profile mostly flagged their messages for harassment, hate speech/threat of violence, and sales or promotion or illegal activities and less likely to flag them for sexual messages/solicitations/nudity and spam/others risk. Overall, this profiles' risk flagging suggested an alignment with their self-reported risk experiences for online harassment perpetration and unwanted online harassment. Understanding the role of the participants in these risks motivated us to uncover the nuances of their direct unsafe conversations, which will be presented in the following section.

4.4.3 The Unsafe Conversations of the Youth Risk Profiles Differed Linguistically (RQ3)

To qualitatively examine the differences that we uncovered statistically in the profiles above, we conducted additional linguistic analyses. Table 4.6 lists the SAGE results as top keywords that were salient in the profiles' unsafe conversations. The following subsections unpack these conversations for each profile.

Low Risks		**Medium Risks**		**Increased Sexting**		**Increased Self-Harm**		**High Risks Perpetration**	
***n*-gram**	**SAGE**	***n*-gram**	**SAGE**	***n*-gram**	**SAGE**	***n*-gram**	**SAGE**	***n*-gram**	**SAGE**
dysphoria	3.86	smelly america	2.60	u wanna come	3.68	cum	3.68	murdered	4.47
brand ambassadors	3.36	respect	2.55	boob pic	2.46	submissive	3.41	die	4.47
kills myself	2.64	hypocritical	2.55	invite you	2.47	nice dick	3.41	burn in hell	4.25
cut it off	2.47	gay	2.51	naked	2.47	mylol	3.41	kick your ass	4.25
blood	2.39	all lives matter	2.49	attached	2.47	eat u out	3.20	stay away	4.25
get rid of	2.32	beautiful	2.48	sleep with you	2.44	jerking off	3.19	weed	4.25
giftcard	2.50	even hot	2.47	airbnb	2.41	vagina	3.19	do not come	4.25
trust	1.81	hoe	2.46	spend the night	2.38	rubbing your pussy	3.19	leave me alone	3.98
your pictures	1.77	are single	2.45	cant be today	2.38	princess	3.13	bitcoin	3.79
hurry	1.76	absolutely gorgeous	2.45	chill	2.38	thigh pics	3.10	where you stay	3.59
someone posted	1.68	thanks friend	2.45	meet you	2.35	i am horny	3.07	vape	3.59
internship	1.60	dumbass	2.44	hang out	2.35	fishnets	3.00	shut	3.51
weekly allowance	1.59	alone	2.43	online school	2.33	lubricant	3.00	little bitch	3.48
wall	1.57	nig	2.42	youre free	2.33	throat	2.98	spyderco	3.32
product reviewer	1.57	ugly	2.42	work	2.29	me so wet	2.97	fortnite	3.15
list name	1.56	stupid	2.41	call or play a game	2.29	leggings	2.97	report you	3.11
sugar	1.56	add snapchat	2.40	today	2.28	your room	2.97	serious danger	3.09
http	1.55	weak ass bitch	2.37	my boobs	2.27	see your pussy	2.96	uncomfortable	3.05
believe	1.55	sexy	2.35	lyk	2.26	ur breast	2.94	kick	2.72
instafans	1.55	cute	2.34	u want	2.26	hurry up bitch	2.94	hurting	2.69

Table 4.6: Top 20 salient *n*-grams (*n*=1,2,3) in the flagged unsafe conversations across the youth risk profiles (SAGE (111)). The *n*-grams with higher SAGE scores are more distinguishing of the profile when compared with the rest of the profiles.

4.4.3.1 Low Risks Profiles: Ignored Spam messages, but Engaged in Scam Conversations and Self-Harm Disclosures of Others

Most of the unsafe conversations of youth that belonged to the Low Risk profile were also flagged as low risk, which mostly included spam, sugar daddy scams, and self-harm disclosures of others. These conversations were distinguished by a variety of spam and scam messages (*brand ambassadors, giftcard, your pictures, someone posted, sugar, weekly allowance, giftcard, instafans etc*), and mental health indicators (*dysphoria, kills myself, cut it off, blood*). The SAGE scores for the self-harm keywords were higher than the spam and scam content, suggesting that other profiles did not have as much self-harm content as the Low Risks profile.

Regarding spam, unsafe conversations contained posted pictures on other accounts or walls (*your pictures*), offers for increasing the number of followers or likes (*instafans*), messages offering gift cards or winning prizes (*giftcard*), or advertisements from business Instagram accounts for their products, services, or their websites. Spam messages were mostly easy for youth to recognize and ignore, but they flagged these messages as risky, as these spam messages made them feel uncomfortable.

The most noticeable type of the scam conversations youth in this profile received was "sugar daddy" requests, where the youth received offers from individuals with promises for high weekly or monthly allowances. During the conversations, although the other individuals usually showed proof of payment, participants refused the offers once they were asked for critical personal information such as credit card or Instagram account credentials. For example, P90 (a 14-year-old female) engaged with someone who promised to send money in exchange for access to her Instagram account, but she soon realized the risk posed by this request and refused.

> **Other user:** *"I'm single 40 years old I have a kid I don't have a wife .. Am looking for an honest baby that will keep my company ... And am ready to spoil her with my money.. can you do this?"*
>
> **P90:** *"of course! could you just send money via cash app? i mean gift card"*
>
> **Other user:** *"Yes baby, but before I send it should I have access on your instagram account for 2 days to gain your trust baby"*
>
> **P90:** *"No don't i trust u"*
>
> **Other user:** *"You don't. If you can let me have access I just wanna buy you gift card"*
>
> **P90:** *"Actually I changed my mind I'm not desperate enough for money to give out my info, Sorry"*

As indicated in the example above, youth appeared to not feel threatened by such advances (i.e., considering them low risk), even trying to take advantage of them, until they assessed risk involved with disclosing their personal information.

Youth in the Low Risk profile also flagged exposure to self-harm as low risk, which was initially surprising to us. After inspecting the unsafe conversations flagged, we found that others (mostly strangers) sought support from the youth by sending private messages, disclosing about their self-harm and/or mental health experiences in their daily lives. For example, a stranger shared their thoughts of self-harm with P158 (19-year-old male) seeking support, which was flagged as low risk self-harm.

> **Other user:** *" i have been having thoughts of wanting to cut i just want to scream and cry for hours"*
> **P158:** *"i'm here if it'd help to talk about things further"*
> **P158:** *"I'm sorry to hear it, self harm urges are awful."*
> **Other user:** *"i can't even do the dishes without picking up a knife to wash it and having the feeling to want to cut right then and there and every night i have been falling asleep crying"*
> **P158:** *"ugh i'm sorry"*

In these cases, youth mostly provided emotional support to others by responding with sympathy. Participants encouraged others to confide in them. Yet, even though these self-harm disclosures were flagged as low risk because participants themselves were not directly in harms way, the fact that they were flagged by youth participants indicates that these conversations still made them (or someone else) feel uncomfortable or unsafe.

4.4.3.2 Medium Risk Profiles: "Cool" Responses to Flirtations and Blowing Off Harassing Comments

Comparing with other profiles, youth who fell within the Medium Risks profile flagged conversations that had a set of distinguished keywords related to sexual flirtations (*beautiful, even hot, absolutely gorgeous, sexy, cute*), personal questions or requests (*are u single, add snapchat*), and harassing comments (*ugly, smelly america, weak ass bitch, stupid, gay*).

Participants in the Medium Risks profile flagged the flirtations they received in Instagram direct messages as harassment, but they usually either ignored these advances or responded positively. For example, when the flirtations came from random strangers (as indicated by participants), like in the following quote from P77, youth rarely responded.

> *"Hi you do not know me but I just wanted to stop by your page to tell you that you're very beautiful ", flagged by P77 15 years old female*

However, when the flirtations came within conversations with non-strangers (i.e., acquaintance, friend, family, or significant other), youth often accepted the compliment with responses, such as *"that's true i'm beautiful, i agree, thank you!"*, emojis that showed they liked these flirtations, or encouragement to send more personal pictures (though not nude). Interestingly, participants still flagged these conversations as harassment, even though their responses appeared as if they welcomed the advances.

Youth in the Medium Risk profile mainly left the harassing comments they received without response, either in one-to-one or in group conversations, yet they flagged these conversations as harassment. Mostly, they received targeted bullying or mean comments about something they had shared publicly. For instance, if youth shared about their sexual identity, they were harassed privately as exemplified in following conversation between P51 (14-year-old, non-binary youth) and another individual.

> **Other user:** *"If u have a set of balls you a dude .. Is it that hard to understand?"*

In other cases, youth were not the target of harassment within group conversations, but they often flagged these conversations as medium risk, reflecting that they felt uncomfortable being involved in a group chat that mainly involved harassment.

4.4.3.3 Increased Sexting Profiles: Flagged Offline Meeting Requests within their Sexual Conversations

Although Increased Sexting and Self-Harm profiles mostly flagged conversations for sexual messages, solicitation, and/or nudity (displayed in Figure 4.3), SAGE resulted keywords from the profiles' unsafe conversations that indicated that youth in the Increased Sexting flagged in-person meeting requests while the youth in the Increased Self-Harm flagged sexual content. Youth within the Increased Sexting profile mostly flagged their conversations for in-person meeting requests resulting in SAGE keywords related to meeting in real life (*u wanna come, invite you, sleep with you, bnb (i.e., Airbnb)*).

While the Increased Sexting profile engaged in sexual conversations, we observed a level of hesitation when the other individuals were willing to transition the relationship into physical meetings sometimes for sex as stated. For instance, in the following conversation, P121 (17-year-old, female) expressed hesitation towards an in-person meeting request from someone within their sexual conversation "in the least sexual way."

> **Other user:** *"My house is free if u wanna come this weekend"*
>
> **Other user:** *"I want to sleep with you (I mean that in the least sexual way possible)"*
>
> **P121:** *"i would but i'm always working on the weekends, but I'll keep it in mind"*

Hesitation from other participants also came in the form of not explicitly refusing these requests; instead, when the other person usually insisted to meet, they sent various excuses to refuse meeting the other person such as being sick, have work, school, or have their family over as well as promising the other person of future meetings.

In addition to the in-person meetings, youth in the Increased Sexting profile flagged other sexual conversations for the requests they received to send their nudes or have sexual video calls. The youth in this profile seemed to only want to be the recipient within these sexual conversations. For instance, while they willingly engaged in such conversations and explicitly agreed to receive nudes from others, many refused sending them and flagged the requests as unsafe, as in the following conversation from P3 (16-year-old, female).

> **Other user:** *"is it ok if i can send nudes"*
>
> **PP3:** *"yeah"*
>
> **Other user:** *"To you?"*
>
> **P3:** *"sure"*
>
> **Other user:** *"ur boob pic?"*
>
> **P3:** *"no"*

4.4.3.4 Increased Self-Harm Profiles: No Self-Harm Content was Found but Engaged in High-Risk Sexual Conversations with Strangers

Comparing with other profiles, Increased Self-Harm, even in comparison with the Increased Sexting profile, had the most sexual SAGE keywords that indicated sexual interactions within their unsafe conversations (*cum, nice dick, jerking off, rubbing my pussy, i am horny, lubricant, me so wet*). Meanwhile, no self-harm content was found in the keywords; however, when inspecting their conversations more closely, we did find indications of mental health struggles, as they flagged the support they received from others as harassment.

Youth in the Increased Self-Harm profile often engaged in sexual exchanges with others who they indicated were strangers. Similar to the below conversation that P82 (17-years-old, female) has, where she explicitly exchanged nudes with a male.

> **Other user:** *"I wanna see your pussi"*
>
> **Other user:** *"Do you wanna see mine??"*
>
> **P82:** *" sure yes"*
>
> **Other user:** *"And can I see yours?"*
>
> **P82:** *"yes"*
>
> **P82:** [user sent attachment] *"looks nice"*

Although youth in this profile stated their enjoyment to sexually please others as in the following example from P94 (16-year-old, non-binary youth) who answered that they enjoyed the experience, they regretted engaging in these conversations based on their reflections (e.g.,"*This conversation was grooming. I regret it.*" from P94), mostly because they discovered an who tried to sexually solicit them. In fact, the Increased Self-Harm profile was the only profile that did not only flag the received messages but also flagged their own messages as unsafe.

> **P94:***You make me so wet.. yes, daddy Fuck me* [Flagged as high-risk sexual messages/solicitation] **Other user:** *"oh girl, u are a good submissive. I fucking love u"*
> **P94:** *"I really want to make you happy daddy"*
> **Other user:** *"awww, daddy is happy"*
> **Other user:** *"Tell me are u enjoying this ? Don't lie"*
> **P94:** *"I am enjoying this daddy"*

Regarding the self-harm content, SAGE did not result any keyword related to sharing any self-harm content within their unsafe conversations. Instead, after we inspected their unsafe conversations, we noticed that the youth in this profile flagged the support they received from people (mostly known) as unsafe. For example, P67 (15-year-old, female) who flagged her friend's (based on her reflection on the conversation) advice as unsafe.

> *"I know Ur Mental Is not as Strong as Mine So...I've told u before Not to trust anyone from MyLOL and Not to trust ANYONE like Literally ANYONE Except Me. "*

This profile seemed to be annoyed by others' protective support in the form of providing unsolicited advice and their stated hesitation on this profile's ability to make correct choices. This profile ignored these messages along with the constant requests of taking care of them and flagged them as unsafe.

4.4.3.5 High Risks Perpetration Profiles: Flagged Threats and Illegal Products

The High Risks Perpetration profile had the highest SAGE scores for the keywords yielded from their unsafe conversations, indicating a completely different risk content from other profiles. This profile's unsafe conversations had serious indications of physical threats and harm (*murdered, die, burn in hill, kik your ass, stay away, etc*) and illegal products (*weed, vape, spyderco*), which were keywords that were not observed in other profiles' conversations.

There were serious indications of threats and bodily harm within the unsafe conversations of the youth in the High Risks Perpetration profile. Most of the conversations appeared to occur after an in-person (mostly at schools or neighbourhood) conflict already happened or started online in community-based game platforms, like Fortnite or communication servers like Discord as stated in the context of the messages then escalated in

Instagram direct conversations, like in the following threats P117 (17-year-old, female) received because of what she did at school.

> **Other user:***"You was talking all that shit at school ho"*
>
> **Other user:***"When you come back I'm going to kick your ass"*
>
> **Other user:** *[participant name] "ass bitch"*
>
> **Other user:** *"Stay away from my nigga"*
>
> **P117:** *"Your white self saying the n word I'm gone report you to the principal Monday"*
>
> **Other user:** *"Bitch shut you ho ass up. You a weak ass bitch"*

The threats ranged from hacking social media accounts to more serious aggressive physical threats of killing or beating. Social exclusion was also observed among the threats received as others threatened this profile of removing them from an online gaming community or staying away from them or their friends in school. In most cases, youth in this profile tried mostly to avoid further escalations by explicitly mentioning that they will report the threats to either authorities or their friends.

The youth in the High Risks Perpetration profile flagged another set of messages for illegal products. We found that youth in this profile engaged with Instagram accounts that sell vape, weed, knifes, or products used for marijuana and sent promotions like *"I have good stuffs for sale, I got weed, pills, research chemicals, carts, vapes, wax , LSD and more."* Unfortunately, the youth in this profile often did not only buy these products easily from such accounts, they also voluntarily reviewed them to others within their private conversations, especially group conversations, such as in the following conversation from P134 (18-year-old, male) who discussed about substance misuse.

> **Other user:***"Big blunt today?"*
>
> **P134:***"smoked a joint.. Im good"*

Overall, in this section, we demonstrated how each of youth risk profiles had distinct characteristics in private conversations that were flagged risky based on their perspective. These findings will be further discussed in the following section.

4.5 Discussion

This study presented five profiles of youth (RQ1): 1) Low Risks (51% of the participants), 2) Medium Risks (29%), 3) Increased Sexting (8%), 4) Increased Self-Harm (8%), and 5) High Risks Perpetration (4%). These profiles' self-reported risk experiences were found to be fairly aligned with their self-assessment of social media trace data (RQ2). We uncovered key linguistic differences in unsafe conversations across the profiles that described each profile's unique risk experiences (RQ3). This section further discusses the implications of these results in comparison to other youth safety literature.

4.5.1 Profiling Youth Risk Behaviors to Develop Targeted Risk Education and Prevention for Youth (RQ1)

Profiling youths' risk experiences (RQ1) uncovered a more nuanced picture of the risks youth encounter both offline and online. Importantly, we found only around 20% of youth experienced the most concerning risks (i.e., Increased Sexting, Self-Harm, and High Risks Perpetration) that intertwined across online and offline contexts and levels of involvement (i.e., victim vs. perpetrator). This finding is significant as it illustrates the multidimensional nature of the risks that youth encounter online and offline. In addition, it shows that most of youth were active social media users, yet only a smaller percentage of this population experienced high level of distinct risks across offline and online contexts. Online risks have been presented by the media as an "epidemic" among youth, adopting a "moral panic" stance that associates youths' social media usage with any moral failings or violent criminality (354). In contrast, our youth profiles showed that this might not be the case for all youth. Therefore, we argue that future risk intervention efforts should focus more on youth who are at high risk similar to the selective prevention strategies that was presented to help at high risk youth such as children of parents with mental illness (24).

Once high-risk sub-populations of youth have been identified, risks prevention programs should design targeted prevention plans catered to the specific types of risk those groups of youth encounter. For example, we found that youth in the increased self-harm reported the highest scores of online sexual risks and offline self-harm while the youth in the High Risks Perpetration profile reported the highest scores for unwanted online harassment and harassment perpetration (refer to Figure 4.1). In contrast, existing cyberbullying interventions that often either target victims (e.g., the U.S. government stopbullying initiative [1]) or perpetrators (e.g., the Centers for Disease Control and Intervention initiative to Reduce Youth Violence (95)) may fail to adequately protect youth who are victims and perpetrators of online harassment. In our results, we observed that sometimes youth experienced the dual experiences of online harassment, which is important to be noted as they could be at a higher risk of reacting to harassment with aggression, depressive, and somatic symptoms (136). The stressful experiences of unwanted online harassment may trigger youth to be perpetrators and have stronger emotions to harm others as a coping mechanism, especially with the anonymity afforded by online spaces (7; 355; 107). Therefore, we recommend future research to design evidence-based interventions that take into consideration both behaviors of youth who are victims of online harassment and may harass others to appropriately help them and mitigate harm.

[1] https://www.stopbullying.gov/

4.5.2 Unpacking the Nuanced Risk Behaviors, Experiences, and Perceptions of Youth (RQ2 and 3)

Our results for RQ2 indicated that the self-reports from youth showed ecological validity and strong correlations with their social media trace data. Therefore, the limitations of self-report data in prior research (276) may not be generalizable across all possible contexts. At the same time, our research demonstrated the value of triangulating youth self-report data with their social media data to uncover key nuances in their lived online risk experiences that would not have been found if we had only focused on the quantitative analysis of survey data. For example, it was unexpected to find youth self-reported self-harm as low risk, which only made sense after we examined their social media data to find that they were acting as supporters of others who brought up self-harming. Disentangling these incidents was important for better understanding why and how youth characterized risks. Therefore, we urge future research to also leverage self-reports of youth backed by their digital trace data to provide more accurate conclusions.

Overall, we found that youth experienced a wide array of risks that characterized the youth profiles' risk experiences in Instagram private conversations (RQ3). Below, we will discuss the important implications of imminent risks to youth such as the physical threats, the nonexistence of self-harm content for the youth who self-reported self-harm, the changes in youths' risk perceptions.

Contradictory to Pabiana's et al. (262) study that suggested that most of youth's offline conflicts or threats remain offline, we found that the youth who belonged to the High Risks Perpetration profile received and flagged threats that stemmed from escalated conflicts as violence/threat of violence/ hate speech. Through unpacking their unsafe conversations, we uncovered the possible reason behind preserving violence/threat of violence/ hate speech as a higher risk than general harassment, which may rely on the fact that these threats indicated purposeful physical harm to the youth. This implication presents an important point for future research to consider separating these risks, which is inline with prior research that highlights the importance of not confiding harassment with hate speech and violence (133). Violence or threats are more targeted to make the person fearful, hate speech is an extreme bias expressed, while harassment is a more of emotional torment (Wachs). In addition, while prior research on youths' violence in the offline context such as school or within family violence is well established (389; 359), less works have investigated online violence. In fact, online violence has been found to lack standard definition and methods (30). Therefore, our findings motivate future research to investigate the youth experiences of online violence, especially that these experiences could present imminent risk to their lives. In this study, we also found that youth who self-reported the highest scores of offline self-harm did not exhibit evidence of digital self-harm. This finding contradicts previous research that found youth who self-harmed often shared self-harm content in social media (64). A possible explanation is that the youth who self-harmed may not manifest these self-harm behaviors online. Instead, they encountered high-risks online sexual interactions. This implication points again to the importance of calibrating self-reported responses with social media data to uncover such nuance. Thus, instead of only relying on social media content to identify youth at-risk for self-harm, which is the case of most of the current literature (229; 123), future research is warranted to consider high-risk sexual behaviors and/or self-reported mental health concerns as a proxy for identifying self-harm risks. Using these as proxies would help scholars, and possibly even clinicians, to more precisely identify the youth who are at-risk of self-harm; and therefore, intervene before physical harm occurs.

Importantly, our findings emphasize on how youth often flagged conversations as unsafe, even when they appeared to engage and enjoy these interactions. For instance, youth in the Medium Risks profile engaged with the flirty comments and youth in the Increased Sexting and Self-Harm profiles engaged in sexual conversations, but then all these profiles flagged these conversations as unsafe. In addition, youth in the Low Risks profile emotionally supported others who disclosed about their self-harm or mental illness experiences, which could reduce the urges of self-harm to others (366), but, could also trigger the imitative self-harm behaviors of the supporters (25; 55). These findings suggest that the youth in these profiles may had a hard time setting healthy boundaries and coping strategies in such situations. Therefore, future risk prevention programs should consider this implication and teach youth how to cope with unwanted situations as well as help them understand how to increase the benefits of supporting others who self-harm and reduce the harmful exposure to these disclosures.

4.5.3 Implications for Design, Education, and Targeted Interventions

Our results highlight the importance of moving beyond a unidimensional view of online risks, towards considering the multidimensionality and interplay between different risk types and settings (online vs. offline). We learned that for certain offline risks, teens may not exhibit the same type of risk indicators in their online behaviors. Targeted risk prevention strategies need to rely on this multidimensionality of risks for improved risk detection and prevention. Apart from risk types, we found that the boundaries between offline and online risks are often blurred (especially for illegal activities), emphasizing the need for a more holistic understanding of online and offline risks together.

Conversely, educational programs that usually teach based on a one-size-fits-all approach for general risk coping and resilience need to evolve to provide specific risk coping education to teens, catered to their risk experiences, which can be vastly different based on their risk profiles. As such, we recommend an integrated approach for education on mitigating both online and offline risk (e.g., bullying) (118), to prevent escalation of risks in both settings. Importantly, the differences in youth profiles did not arise based on their personal characteristics, but were rooted in their lived risk experiences, highlighting the importance of profiling youth based on their unique experiences, rather than their personal traits which may result in harmful racial or gendered profiling (302; 293). Additionally, creating profiles of youth based on their self-reported risk experiences demonstrated the fair alignment of these self-reports with social media trace data. Therefore, we encourage researchers, social media platforms, and practitioners to leverage both types of data to provide more reliable machine learning models to detect and mitigate risks targeted for youth profiles. For instance, these machine learning models could be tailored to the multidimensionality of youths' risks by jointly considering risks such as sexual and self-harm (similar to the Increased Self-Harm profiles), which would help target youth who are experiencing multiple high risks and in most need of intervention (e.g., sexual and mental health resources).

Further, current Artificial Intelligent (AI) risk detection solutions were developed and marketed without public evaluation (301), especially from youth. Incorporating youth evaluations when building AI risk detection solutions would not only enhance the accuracy of the AI models, but also ensure digital equity, especially for socio-economically disadvantaged youth (289). An optimal way to move towards designing youth-centric AI solutions is by leveraging the human-in-the-loop approach (239) to use youth assessments for the AI risk predictions to inform about the quality of the risk predictions, which could further enhance the accuracy of the models. At the same time, being able to accurately flag and reflect over these past experiences may help youth develop a level of risk self-awareness, which would benefit their future online communication through reflective learning. Moreover, instead of gaining awareness from regrettable past interactions, we encourage helping youth to be safer in real time by providing intelligent assistance (e.g., nudges) (222) to teens that prompt them towards safe responses. These real-time self-assessments would equip them with the necessary skills, resilience, and awareness for navigating unavoidable risks in the future.

4.5.4 Limitations and Future Work

While collecting Instagram private conversations from youth is a key strength of this study, it may also affect the generalizability of our results. Since our results were based on youth experiences on Instagram, they might not be generalizable to other social media platforms characterized by different youth demographics, moderation strategies, and/or affordances. Therefore, we recommend future research to investigate risks that occur on other platforms to validate the alignment between self-reports and digital trace data. Furthermore, this study was conducted with youth (ages 13-21) in the United States; therefore, the results should be generalized to only this youth population. Future research is warranted to conduct the study across different countries, where the GDPR rules might be applied as well as the different cultural norms that may influence youth offline and online behaviors. The participants were self-selected to participate in this study, which investigated the risk behaviors of youth. This sample could be subject to sample bias toward the ones who are victims of online risks or abuse. While we indeed found a large portion of our participants identified themselves as LGBTQ, our youth risk profiles indicated that over half of the participants experienced low to medium self-flagged online risks. Therefore, we argue that the sample that participated in this study is a fairly representative sample of youth in the United States.

CHAPTER 5

STUDY 3: Timeliness Matters: Leveraging Reinforcement Learning on Social Media Data to Prioritize High-Risk Conversations for Promoting Youth Online Safety

Ensuring the online safety of youth has motivated research towards the development of machine learning (ML) methods capable of accurately detecting social media risks after-the-fact. However, for these detection models to be effective, they must proactively identify high-risk scenarios (e.g., sexual solicitations, cyberbullying) to mitigate harm. This 'real-time' responsiveness is a recognized challenge within the risk detection literature. Therefore, this paper presents a novel two-level framework that first uses reinforcement learning to identify conversation stop points to prioritize messages for evaluation. Then, we optimize state-of-the-art deep learning models to accurately categorize risk priority (low, high). We apply this framework to a time-based simulation using a rich dataset of 23K private conversations with over 7 million messages donated by 194 youth (ages 13-21). We conducted an experiment comparing our new approach to a traditional conversation-level baseline. We found that the timeliness of conversations significantly improved from over 2 hours to approximately 16 minutes with only a slight reduction in accuracy (0.88 to 0.84). This study advances real-time detection approaches for social media data and provides a benchmark for future training reinforcement learning that prioritizes the timeliness of classifying high-risk conversations.

5.1 Introduction

Approximately 90% of U.S. youth between the ages of 13 and 17 have a social media account, with platforms like Instagram being used nearly daily (349; 188). Although this widespread usage offers youth the chance to acquire knowledge, explore identity, and interact with others, it also opens them up to online dangers, like sexual solicitations and cyberbullying, as well as mental health risks that can manifest both on and offline (248; 292; 261). The adverse impacts of these online-offline risk interactions can be long-term, significantly affecting teens' mental health, self-esteem, and overall well-being (248). Therefore, Social Computing and Human-Centered Machine Learning (HCML) researchers have highlighted the urgency for effective ways to identify these risks, as well as interventions tailored to youth to proactively protect them online (65; 167). Significant strides have been made that reflect a shift in focus from technology alone to a more holistic view that includes human elements in the design and implementation of risk detection models (290; 179). Nevertheless, there is still a critical area for improvement, particularly in existing training methods that detect the risks without accounting for timing, highlighting the need for enhanced real-time risk detection to ensure timely and immediate response to potential threats.

In particular, advancements in risk detection need to be *responsive* to the rapid evolution of online interactions and how online risk manifests, necessitating immediate or real-time interventions that are context-specific to reduce harm, particularly among youth (22). First, the current reliance on entire conversations might be inefficient for timely risk detection, while the reliance on single messages could be insufficient for a comprehensive understanding of context (305). This limitation highlights the need for more accurate context-sensitive analyses to identify risks that may be overlooked when only considering individual elements or the entirety of conversations (286). Secondly, traditional methods that evaluate online conversations for risk detection without differentiating between their varying levels of risk severity often struggle to achieve responsiveness. Thus, it is crucial to adopt a triage approach that focuses on identifying interactions that pose higher risks by directing the right type of attention to those who need it the most (18).

Lastly and most importantly, for risk detection models to be effective in the dynamic realm of online interactions, their evaluation must extend beyond accuracy to include *timeliness* (312). Timeliness as a measure is often evaluated based on the time elapsed between the initial occurrence of a potentially harmful interaction and the moment a response or intervention is initiated (350). The shorter this duration, the more timely the response is considered, indicating a high efficiency in recognizing and addressing risks promptly. Thus, the ability of a model to quickly prioritize high-risk conversations or detect potential risks is as important as its accuracy in identifying them because harm can significantly increase over time if the sources of risk are not addressed quickly. To this end, in this paper, we present a novel real-time two-level algorithmic framework for prioritizing high-risk conversations within youth online interactions. To fulfill this objective, we address the following research questions:

- **RQ1:** *How can we identify optimal stopping points for evaluating conversation level of risk priority?*
- **RQ2:** *How can we optimize deep learning models to accurately prioritize high-risk conversations?*
- **RQ3:** *How is timeliness impacted by the prioritization approaches?*

To answer these research questions, we obtained access to the Instagram Data Donation (IGDD) dataset collected by Razi et. al 2022b, which contains 23K risk-flagged private conversations with over 7 million messages of youth (aged 13-21) on Instagram. We leveraged youths' ground truth risk annotations of their risk experiences to train classifiers used for the real-time conversation prioritization algorithm.

In this study, we developed a two-level algorithmic framework. The first level involves training a reinforcement learning agent to determine the appropriate point to stop reading a conversation's messages and forward them for evaluation to answer RQ1. At the second level, a deep learning model assesses whether the conversation is of high or low priority based on the risk severity in the messages to address RQ2. The primary contribution of this paper is on the novel approach for conversation prioritization; therefore, we integrated pre-existing and pre-trained models for risk detection from published works by (14; 286; 15) to round out our framework's capabilities. This includes models for identifying sexual content and cyberbullying. For RQ3, we evaluated our proposed real-time conversation prioritization framework with two key goals: firstly, to demonstrate the efficacy of our RL agent decisions of the evaluation points in contrast to a baseline approach that processes conversations in their entirety; and secondly, to compare the overall processing time required for assessing all conversations against an approach that concentrates exclusively on conversations deemed high-risk.

In this paper, we presented a benchmark for training reinforcement learning algorithms to identify the optimal moment for halting and assessing messages in the context of conversation priority detection. Additionally, we found that identifying conversation priority at the conversation level achieved superior accuracy (Acc. 0.88) compared to an individual message level (Acc. 0.82). However, this method could potentially delay the allocation of resources. To address potential delay, we showcased the integration of a benchmarked reinforcement learning model with evaluation points, effectively balancing the determination of when to send a conversation for a priority check. Our model, which included evaluation points for prioritization, achieved a higher accuracy rate (Acc. 0.84) than the message-level model (Acc. 0.82), but lower than the conversation-level model (Acc. 0.88). it offered faster detection responses than waiting for the complete conversation. Therefore, through this work, we shed light on the empirical and practical implications of efficient conversation prioritization and the balance between accuracy and timeliness in determining conversation priority. In summary, this work makes the following novel research contributions:

- We created a novel two-level framework with the ability to identify high-risk conversations, which represents a significant advance in the field of real-time risk detection for prompt risk mitigation of high-risk online conversations of youth.
- We establish a benchmark for training a reinforcement learning agent in the context of dynamic online environments, setting a foundation for future advancements that would broaden the scope and effectiveness of real-time risk detection mechanisms.
- We demonstrate the need for balanced solutions that effectively manage the dual requirements of precision and quick response in risk detection by empirically illustrating the trade-offs between accuracy and timely identification of conversation priority.

Next, we will synthesize the related work that motivated the need for the creation of our conversation priority framework.

5.2 Background

5.2.1 Automated Risk Detection for Youth

Research on online risk detection for youth is well-established and marked by significant advancements. Numerous experts in machine learning and computational social science have developed and evaluated automated detection algorithms to detect various risks in social media such as sexual risks, cyberbullying, and mental illness (17; 256; 225; 14). In prior work, scholars have mainly aimed to create and present accurate automated detection models for these risks by heavily relying on traditional machine learning approaches such as supervised and semi-supervised models and ensemble methods. For instance, Ali et al. (14) leveraged an ensemble approach to detect unsafe conversations of youth and showed that classifiers that were trained based on metadata and relationship types performed better in terms of classifying conversations as safe or unsafe, with 87% accuracy. As the field matured, limitations have emerged, notably how these detection algorithms could overlook the nuances of human interactions and fail to adapt to the evolving nature of online risks. This led to a paradigm shift towards HCML, an approach that places humans, youth in our case, at the core of algorithm design and development (290; 179; 66).

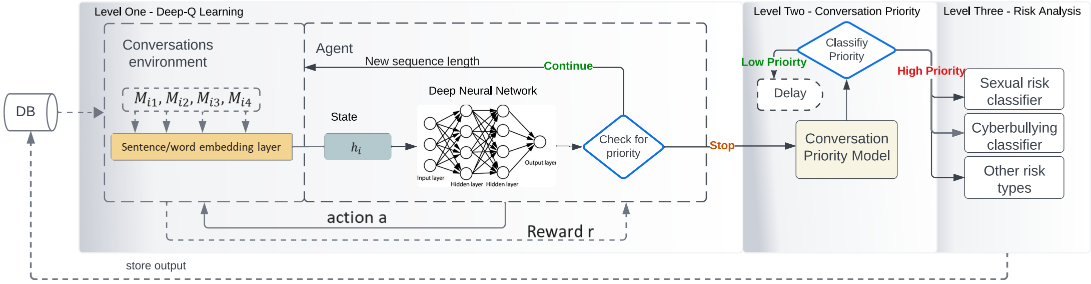


Figure 5.1: Structure of the conversation priority framework. The trained agent determines the optimal moment to route a specific conversation for priority evaluation. Conversations labeled as low-priority were deferred for risk analysis, whereas those identified as high-risk were directed to a dedicated high-priority pipeline for more in-depth analysis by the risk detection algorithms within the third level that is out of the scope for this paper.

A crucial element in creating human or youth-centered risk detection models is highly dependent on the quality of the dataset utilized in their training (179). Consequently, HCML scholars have placed a significant emphasis on establishing ecologically valid, ethical, and trauma-informed data collection practices designed precisely for vulnerable populations like youth. The Instagram Data Donation project (IGDD) (287) represents a significant recent effort in this area, focusing on a human-centered methodology to gather and analyze private social media conversations from young individuals aged 13 to 21 years to advance risk classification research. This initiative highlights both ethical challenges (e.g., reporting child abuse for an immediate risk to minors) and technical hurdles (e.g., enhancing system efficiency for the upload of Instagram data files), all pivotal in the collection of such challenging-to-access datasets. Given the sensitive nature of the data involved, the researchers prioritized protecting the privacy of participants by restricting access to this sensitive data and allowing data sharing only through collaborative partnerships under a data usage agreement that legally protects the data from unauthorized access and/or misuse. Through these collaborative efforts, this rich dataset has been distributed among researchers, allowing them to build automated models to detect risks targeted to youth based on a human-centered understanding of their behaviors and risk perspectives (14; 286; 270; 13). For example, Park et al. (270) used this dataset to conduct qualitative content analyses on media content flagged as risky by youth to inform the development of semi- and self-supervised vision transformers for media risk detection. The study found that vision transformers were able to effectively learn complex image features for automated detection of contextualized media labels (e.g., harassment, screenshot, and personally targeted). Although such initiatives have been beneficial in developing human-centered algorithmic methods, current models were geared towards recognizing risks post-occurrence, instead of using advanced techniques to proactively identify potential risks early enough to prevent victimization or minimize the damage. Therefore, we address this gap by focusing on providing a framework that would enhance the timeliness of risk detection models by utilizing the IGDD data.

5.2.2 Towards Real-Time Risk Detection

In recent research, scholars have highlighted the need for more timely approaches in detecting and addressing online risksby examining time-sensitive risk detection across various issues such as depression, self-harm, pathological gambling, and eating disorders using social media datasets (267; 266; 87). The majority of the research in real-time risk detection relied on deep learning models such as Convolutional Neural Networks (CNN) and Long Short Term Memory (LSTM), particularly suited for the sequential understanding and representation of data. For instance, Liu and Wu (206) used CNN for real-time fake news detection by adding an attention mechanism, which was used to learn how much attention should be given to the data points in the sequence. The essence of such solutions was captured in performance-based evaluations, notably, accuracy performance over time (101), underlining the importance of timely responses in risk mitigation strategies.

Yet, a key computational challenge has been acknowledged, where real-time models struggle with the vast scale of social media data, impacting their responsiveness and effectiveness (282). To address this challenge, a few studies have introduced a method for prioritizing conversations that would invoke risky interactions (282; 382). Reducing the number of conversations or messages that need to be classified or queued for risk assessment decreases the computational burden associated with feature calculation, which in turn, leads to quicker decision-making in risk detection (382). For instance, Rafiq et al. (282) introduced a Dynamic Priority Scheduler (DPS) that actively allocated high priority to Instagram sessions (i.e., posts and their comments) requiring immediate review by the detection model, while assigning lower priority to others that can be deferred until new comments emerged.

While this scheduling approach demonstrated superior responsiveness compared to conventional methods, there is still a need for adaptive and more generalized allocation approaches that allocate resources based on the volume of conversations and the urgency of risk detection. One way to achieve this is by Reinforcement Learning(RL) (33) that could learn when potentially risky conversations would need the detection algorithms. Building beyond this research, this paper will be the first that leverages reinforcement learning to create a novel algorithmic framework that focuses on high-risk conversations, as illustrated in Figure 5.1. This framework is structured around two key levels: 1) An agent responsible for determining whether to route a specific conversation for priority evaluation or to maintain it within the standard conversation stream, and 2) In scenarios where the conversation is flagged for priority check, the second level, a pre-trained deep learning model, assesses the conversation's priority level. High-priority conversations are shifted to a specialized high-priority pipeline, while low-priority ones are tagged accordingly and retained within the normal conversation flow. To complete the algorithmic cycle, advanced risk detection algorithms conduct a thorough analysis of the conversation's content within the high-priority conversations pipeline, ensuring continuous monitoring and evaluation of these high-priority interactions.

5.3 Methods

5.3.1 Instagram Data Donation Dataset

In this work, we utilized data from a youth Instagram Data Donation (IGDD) project by Razi et al. (287), which was approved by their Institutional Review Board (IRB). Instagram is one of the most popular platforms among young individuals (188), which was why we selected the dataset.

The dataset included Instagram Direct Messages (DMs) from 194 U.S.-based English-speaking participants aged 13-21. These individuals had active Instagram accounts for over three months during their teens (13-17 years), interacted through DMs with at least 15 people, and had a minimum of two uncomfortable or unsafe DM conversations. Participants downloaded and uploaded their Instagram data and flagged their conversations as safe or unsafe. They also categorized unsafe messages according to risk types (i.e. sexual solicitations, harassment, etc.) derived from Instagram's reporting feature risk categories [1] and risk levels (i.e., high, medium, and low) grounded from existing research (362). **Low Risk** messages were those that caused discomfort to the participant but were not likely to lead to emotional or physical harm. **Medium Risk** encompassed messages that had the potential to cause emotional or physical harm if they were to continue or escalate. **High Risk** messages were identified as dangerous, having already caused emotional or physical

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

harm to the participant.

5.3.1.1 Characteristics of Dataset and Participants.

The dataset was gathered from 194 young individuals aged 13 to 21 years, with the mean age being 17 years and a standard deviation of 2.21. Among them, 68% identified as female, 22% as male, and 10% as either non-binary or did not specify their gender identity. The dataset contained a total of 23,089 Instagram private conversations, with a range from a minimum of 17 to a maximum of 1477 conversations. These conversations encompassed over 7 million messages. Participants labeled 2,760 of these conversations as unsafe, causing feelings of discomfort or unsafety. These unsafe conversations contained 205,187 messages. Out of these messages, 3642 messages were labeled by participants for risk levels, with 2,025 being labeled as low risk, 1,127 as medium risk, and 490 as high risk. In this paper, our emphasis was on conversation priority for the triage approach, which was dependent on the levels of risk severity. Consequently, we limited our analysis to these risk levels, reserving the examination of specific risk types for a forthcoming paper.

5.3.2 Data Processing

In the data preprocessing stage, we eliminated punctuation, hyperlinks, stop words, non-Latin words, single or numeric characters, and conversations with less than three words. To retain the semantic value conveyed by emojis, we converted them into their word equivalents using the 'demoji' Python library [2]. This process was critical to maintaining the meaningful content of the conversations while filtering out irrelevant elements.

Our training approach for both the conversation prioritization and the RL model utilized the unsafe conversations only with its 205,187 messages. Messages lacking risk labels were classified as safe. For training, risk level labels were numerically encoded (safe:0, low:1, medium:2, high:3). We merged the 'low' and 'safe' labels into a '0' label to indicate low-level risk, and combined 'medium' and 'high-risk' levels into a '1' label for high-level risk messages. This decision was made because the actions for both safe and low-risk level messages would be the same, involving a delay in the evaluation of messages. Conversely, messages labeled as medium to high risk required immediate attention by the risk classifiers. Therefore, combining these categories streamlined the response process, aligning actions with the appropriate level of urgency needed for each risk category. In our approach, we focused on the textual features of the messages to establish a benchmark for training the reinforcement learning algorithm by demonstrating the algorithm's effectiveness using basic input, without the need to create complex features. The dataset was split 80% for training and 20% for testing.

We conducted experiments with transfer learning using Google Research's BERT-based ELECTRA model

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

and word2vec. It was observed that using a pre-trained ELECTRA model, which had been trained on general English language data for generating message-level sequence embeddings, improved the classification accuracy more than the word2vec embedding. Due to memory constraints, we utilized the HuggingFace Transformers library for an ELECTRA-based implementation but could not deploy the larger model variant. The ELECTRA model generates a sequence embedding for each input message that represents the words in the message within the context of other words in the same message. Through transfer learning, we leveraged the outputs of these models as inputs for our RL and conversation prioritization model.

5.3.3 Models

5.3.3.1 Two-Level Conversation Prioritization Algorithmic Framework

The objective of early conversation prioritization was to identify the conversation priority as high or low based on the likelihood of having unsafe messages as early as possible while keeping an acceptable accuracy performance. We employed three models: 1) an RL model that identified the stopping point for the model to review messages and trigger the prioritization model, 2) a deep learning model that classified the conversation priority, and 3) pre-trained risk detection algorithms. Unlike Rafiq et al.'s work (282), we did not rely on a threshold to set the priority category for the conversations as this threshold was identified based on their dataset which lacked generalizability to other datasets. Instead, we relied on first-persons' (i.e., youth) perspective of the risk severity as ground truth, which could be considered as a more ecologically valid and human-centered practice to rank the conversation priority (290).

5.3.3.2 Reinforcement Learning Model-Level One (RQ1).

We utilized an RL model to determine the point at which messages needed to be evaluated for priority. The RL model was designed to reward based on the actual priority labels of the messages provided by the youth and apply a minor penalty when it failed to trigger the classifier. This approach allowed our prioritization algorithm framework to learn the balance between accuracy and timeliness in decision-making.

For the RL model, we chose the Deep Q-Network (DQN) (238). A significant hurdle in Q-learning involves finding a reliable action-value approximation, enabling precise mapping from (state-action) pairs to expected returns in real time. This becomes particularly challenging when dealing with nonlinear system dynamics or reward patterns, similar to our case as the conversations can have varying lengths and content, making the state space complex, and the relationship between the number of messages in a conversation and its prioritization to be nonlinear and intricate. Therefore, the DQN, a deep reinforcement learning method, was the best option to be used because of its ability to handle high-dimensional state representations and its ability to model complex and nonlinear relationships, which may not be achievable with simpler and off-policy algorithms.

The training of DQN was done using kerasRL2 [3], and TensorFlow 2 (1) with integration of Gym (53) to build our custom environment. The action space consists of two options: 0, signifying the action of continuing to read messages without assessment, and 1, signifying the action of halting to forward the messages that have been previously read and trigger the conversation prioritization model. For a given step t, the agent reads the messages that were sent from the environment and takes action. Once the agent made an action for the given step, the environment calculated the reward. The reward was determined based on the risk severity levels labels when the agent took an action whether it was continuing to read messages or stopping for evaluation, where r_i can be expressed as follows:

$$r_i = \begin{cases} \varepsilon & \text{if correct continue,} \\ -\varepsilon & \text{if incorrect continue,} \\ \log(M) & \text{if correct stop,} \\ -P & \text{if incorrect stop.} \end{cases} \tag{5.1}$$

If it was a *continue* action and the risk levels in the messages were either safe or low risk, the ε was a minor reward applied since the data was imbalanced and most of the messages were either safe or low. $-\varepsilon$ was a delayed penalty for having high-risk messages while choosing to delay the evaluation. If the decision was *stop* for evaluation and there were high-risk messages, then $\log(M)$ was a reward where M was the number of correct stops have been made so far. When the decision was *stop* and the messages were low-risk or safe messages, a penalty of $-P$ was applied since the agent asked for the evaluation model to review the messages when it was unneeded.

DQN leverages the adherence of the optimal action-value function $Q^*(s,a)$ to the Bellman equation (331) to update the neural network weights θ_i for minimizing a specific loss function during each iteration i, which occurs after each time step. Once the DQN was trained, and $Q_\theta(s,a)$ was obtained, the action taken by the agent was determined as follows:

$$a(s) = \arg\max_a Q_\theta(s,a)$$

This equation calculates the action $a(s)$ that maximizes the estimated action-value function $Q_\theta(s,a)$.

5.3.3.3 Conversation Prioritization Model-Level Two (RQ2).

Our approach involved training various deep learning models to determine the most effective one for conversation prioritization. This included LSTM, BI-LSTM, CNN, and Deep Neural Network (DNN). The

[3] https://github.com/inarikami/keras-rl2

architecture of our models was broadly similar, incorporating layers such as GlobalPooling on ELECTRA's sequence outputs, a network comprising 5-6 layers along with a layer of the previously mentioned models like a CNN layer, followed by a Dropout layer, and a sigmoid activation classification neuron. This final neuron generated a value ranging from 0 to 1, where 1 indicated a high-priority conversation and 0 indicated a low-priority one.

Both conversation prioritization and DQN models were trained jointly, where the DQN model generated a sequence of instances and the conversation prioritization model acted as the discriminator. The training process was done alternatingly, where training one model and keeping the parameter for the other model fixed and vice versa. The model converges when the reward value stabilizes between consecutive episodes.

5.3.4 Timeliness Assessments (RQ3)

To determine the performance of our conversation prioritization model, we compared its response time to a model that treated all conversations equally through a simulated scenario employing both approaches the baseline and the prioritization model. Additionally, we employed the time elapsed between the timestamp of the message that activates the prioritization model and the timestamp of the high/medium risk level message. Using these metrics, we gauged the lag between the occurrence of the high/medium risk message and the point at which the trigger was activated to evaluate the conversation.

Risk	**Prec.**	**Rec.**	**F1**	**AUC**	**Acc.**
Cyberbullying	0.79±0.02	0.79±0.04	0.79±0.03	0.80±0.13	0.82±0.05
Sexual Risk	0.85±0.02	0.88±0.02	0.86±0.02		

Table 5.1: The accuracy Performances of the Risk Detection Classifier from prior works.

To understand the impact of the conversation prioritization approach versus treating all conversations equally, we ran a simple simulation using a Python code where we sent API requests to an AWS server (m5-large) that hosted pre-trained models that were trained using the same dataset adopted from prior works, including (14; 286; 15) to detect risks such as sexual and cyberbullying, with detection performances listed in Table 5.1. Then, our system recorded the total processing time the server took to process these conversations. We randomly sampled 20K conversations from the dataset that were not part of the training process. Then, we ran the trained conversation prioritization model (with the stopping point) against them, resulting in 1,875 predicted high-priority conversations. Through this test, we were able to compare the total processing time when the system had a prioritization approach versus when treating all conversations equally. In the next section, we present the results of this paper.

5.4 Results

5.4.1 Reinforcement Learning Model (RQ1)

The agent's training was done over a set of 16,000 episodes. Figure 5.2, provides insights into the average reward for episodes. Notably, the reward values started at around -900, which was a significantly high negative value. This suggests that the initial strategy or actions chosen by the DQN were far from optimal, which means it initially stopped reading the messages within a given conversation too early or too late, resulting in negative reinforcement. Then, the rewards increased (become less negative) to around -100 to -250. This improvement indicates that the DQN began to learn a more effective strategy for deciding when to stop and evaluate messages within a conversation. The DQN's actions started aligning better with the optimal points for stopping, as evidenced by the reduced penalty. We also found that the fluctuation in the reward values around -100 to -250 signifies that the DQN was fine-tuning its strategy.

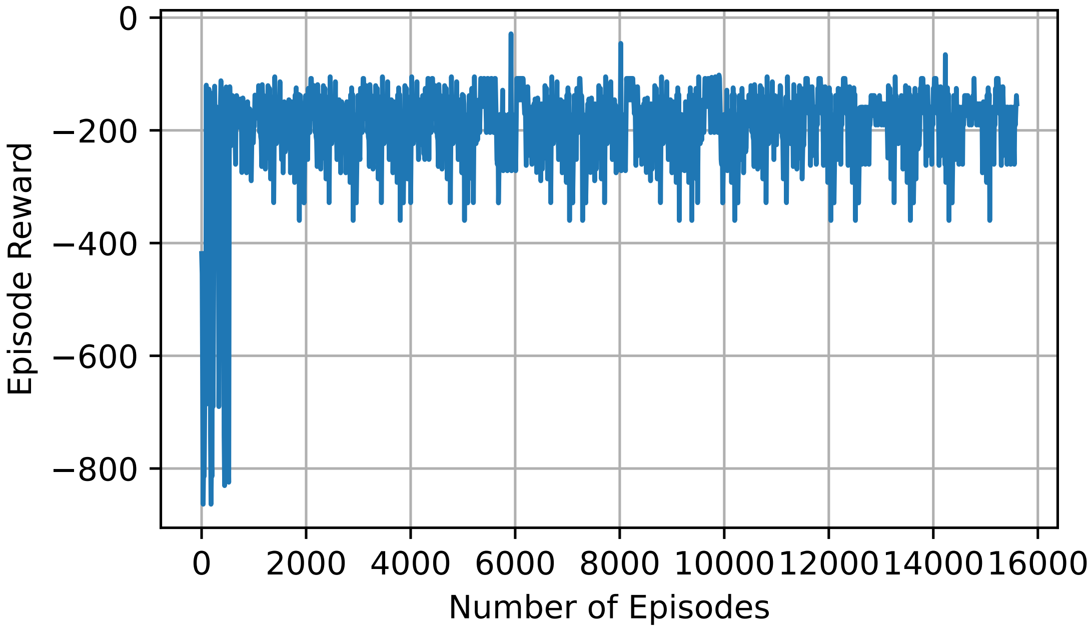


Figure 5.2: Average reward over episodes.

5.4.2 Conversation Prioritization (RQ2)

In this section, we reported the evaluation of the accuracy of models in classifying conversation priorities. While all baseline models and benchmark models classify the conversation priorities using all messages from the conversation, Razi-CNN (286) performed the best on classifying the risk severity levels at the individual message level, using the same IGDD dataset. Additionally, for our presented model, we illustrated two models' performances: (1) our model, which classified the conversations using a selected subset of messages determined by the DQN, and (2) the baseline model, which utilized the entire set of messages within a conversation for the classification of conversation priority. Table 5.2 summarizes the performance metrics of the models on the dataset.

Model	**Prec.**	**Rec.**	**F1**	**AUC**	**Accr.**
Razi-CNN (Messages Level)*	0.82	0.82	0.82	**0.88**	0.82
Bi-LSTM (Conversation level)	0.88	**0.88**	**0.88**	**0.88**	**0.89**
LSTM (Conversation level)	0.85	0.86	0.85	0.85	0.85
CNN (Conversation level)	0.84	0.84	0.84	0.84	0.84
DNN (Conversation level)	0.82	0.81	0.81	0.81	0.82
Bi-LSTM (Stopping Point)	0.86	0.82	0.84	0.83	0.84

Table 5.2: Accuracy performances comparison of our state–of–the–art conversation prioritization model. '*' denotes values taken from the original publications.

Overall, having the entire conversation as input improved the classification of the risk levels as we found that the Bi-LSTM model at the conversation level outperformed Razi-CNN at the message level, which implies that Additionally, this model outperformed all other state-of-the-art models at the conversation level, showing a high level of performance with the best recall and F1 score at 0.88, and the highest accuracy at 0.89. Therefore, the Bi-LSTM model was then deployed with the DQN model, so the input was based on the stopping point, yielding an F1 score of 0.84. There was a marginal increase in the precision of the Bi-LSTM model when applied at the stopping point with a Precision of 0.86, compared with the recall and other metrics experienced a decrease in performance, with the AUC and accuracy both reporting at 0.83 and 0.84, respectively.

5.4.3 Evaluation of Timeliness Performances (RQ3)

While our prioritization model determines the stopping point dynamically, we compared the timeliness measures of this model against the baseline model that required the availability of all messages, which resulted in the best accuracy performance.

We found notable differences in the BI-LSTM model with the stopping point and with the conversation level performance in terms of timeliness as shown in Figure 5.3. The baseline model that processed entire conversations demonstrated a markedly different distribution. Its values were less concentrated around zero difference and tended more towards positive values. The positive skew in this model's distribution illustrated a delay gap between the occurrence of high-risk-level messages and the end of the conversations of between 0 to approximately 5K minutes (i.e., 3 and half days), showing its tendency to identify the high-risk conversation later. BI-LSTM with the stopping point showed a distribution predominantly centered around zero difference

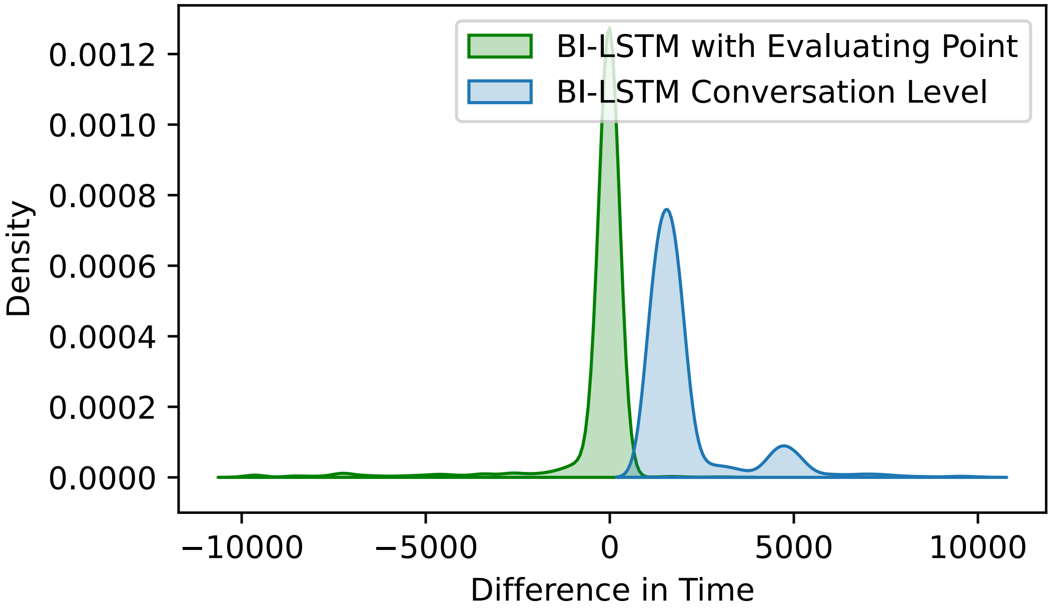


Figure 5.3: Timeliness comparison between stopping point model and the best-performing baseline at the conversation level

in the time elapsed between the message that triggered the stopping action and the timestamp of the risk messages with a slight shift towards negative values. This shift indicates the presence of risk messages occurring after the model decided to stop, in this case around 16 hours before the risks occurred.

Figure 5.4 shows the time processing differences between our conversation prioritization approach and the approach that treats all conversations equally. In the prioritization system, a significant concentration of conversations was found within a time frame of 0 to 16 minutes, roughly equating to 1 million milliseconds. This means most high-priority conversations were processed faster, with the entire batch of high-risk priority conversations being processed in the server within a 39-minute window.

Conversely, when the system treated all conversations equally, a different pattern was shown on the density graph. We found a lower intensity observed across most of the graph, suggesting a more even distribution of conversation processing times without any significant peaks. However, a notable spike in density was found over the two-hour mark (over 5 million milliseconds), indicating a concentration of conversations being processed around this duration. This peak reveals a slower overall processing time, with the system taking close to 2.5 hours to process the conversations.

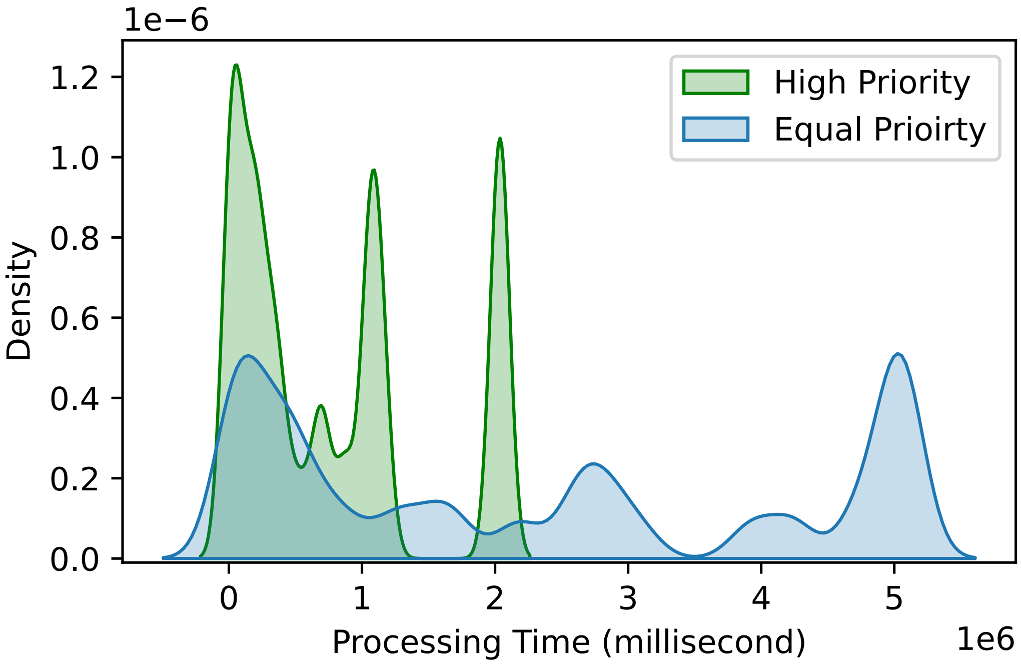


Figure 5.4: Density Distribution of Processing Times for Models with Conversation Priority versus with Equal Priority.

5.5 Discussion

5.5.1 Considerations for Risk Detection Research

Our work highlights that by adopting conversation prioritization methods, timely and efficient decision-making for high-risk conversations is achievable, utilizing less time and fewer resources compared to the analysis of entire conversations. While not all social media interactions need to be monitored equally (282), current practices in the real-time risk detection literature focus on classifying all social media conversations without having a process to prioritize high-risk conversations. Most data processing approaches in prior research include segmenting the dataset into chunks of data of equal size, predefined by a fixed time window (193; 213) or by a fixed number of posts (190; 383), to be fed to the training models sequentially, which are not representative of real-world peoples' interactions. Our work was the first to apply RL and deep learning approaches to optimize conversation prioritization for real-time risk detection with acceptable accuracy performance (0.84 accuracy). Our results confirmed that with RL, the model finetunes itself to find optimal points to identify priority evaluation points within conversations. With optimal evaluation points, we showed that our deep-learning model can significantly increase the timeliness of the detection, which is critical for providing timely and customized interventions to mitigate high-risk interactions on social media. Therefore, instead of relying on pre-defined chunks of data without consideration of risk priorities, we call for future research to apply our RL approaches and the algorithmic framework as benchmarks so that the real-time risk detection models can respond to high-risk social media interaction promptly.

The observed learning curve of our RL agent within the conversation environment, characterized by initial high negative rewards followed by a gradual increase and subsequent fluctuation, is indicative of a typical reinforcement learning process (236). This pattern highlights the significant impact of the structure of our reward function on the agent's learning trajectory, which was designed to balance various factors: minor rewards for 'continue' actions in low-risk scenarios due to data imbalance, penalties for delaying evaluation in the presence of high-risk messages, and substantial rewards for stopping for evaluation at appropriate junctures, particularly in high-risk situations. This nuanced approach to reward structuring could guide future research in real-time risk detection as it highlights the importance of carefully considering the reward mechanism to align with the environments characterized by data imbalances and levels of risk. Therefore, future research is encouraged to explore deeply how different reward function structures could optimize learning for real-time conversation prioritization, especially in environments with similar complexities. Understanding these dynamics could lead to the development of more sophisticated RL models that are better suited for the nuanced and dynamic nature of real-time risk detection in social media platforms.

Another key contribution of this paper is that we empirically illustrated an important dichotomy between achieving high accuracy and attaining fast response times in prioritizing conversations. We observed that when the entire conversation was processed by the model, it achieved higher accuracy in identifying the conversation priority. However, this comprehensive analysis came at the cost of speed, leading to slower prioritization of high-risk conversations. Conversely, our prioritization model with the stopping points, designed to identify the conversation priority earlier in the conversation, demonstrated a marked increase in response speed, with a slight reduction in accuracy. While many scholars have called for timely and accurate risk detection algorithms (60), we highlight that this balance could be difficult to achieve because accuracy forces the full context (messages) to be available while timeliness forces the work with an incomplete context. Therefore, future research, especially within the youth online safety area, is urged to identify the cases where accuracy could be prioritized over timeliness and vice versa.

Specifically, online youth safety literature needs to identify the cases where a rapid response may be essential to prevent immediate harm, while in others, a more accurate understanding of the context might be needed to avoid false alarms or inappropriate actions. Whether to prioritize the slower conversation-focused model or the faster dynamically chosen subconversation-focused model could depend on what type of intervention we seek to provide. For instance, if the goal of the intervention is content moderation or shadow banning, slower models with more caution are warranted as wrongfully banning accounts, removing content, or hiding content for certain audiences could adversely impact user experience and the notion of freedom of speech (187; 342). On the other hand, if the goal is to provide immediate support to potential victims, then faster models would be necessary. For instance, faster sub-conversation-focused risk detection algorithms can be used to trigger real-time nudges or warnings for potential victims of harassment messages or for potential perpetrators, before they send out offensive content (4). Even if the model makes inaccurate decisions (e.g., false positives), such proactive intervention approaches would still be more beneficial than slower models given the persistent trauma that the victims can experience (185; 74).

5.5.2 Considerations for Real-World Implementation

Our simulation results showcased the enhanced responsiveness of leveraging a conversation prioritization technique that focused on conversations with potentially high-risk messages over treating all conversations equally. Content moderation in online platforms is a resource-intensive task, often limited by constraints such as time and workforce availability (164). Thus, implementing an even imperfect system of conversation prioritization could alleviate the burden on human moderators by streamlining the process and focusing their attention on the most critical cases first. Yet, implementing this approach in the real world requires careful consideration. The incorrect prioritization of conversations by an algorithm could have several significant consequences, The consequences of such prioritization errors may vary, including delayed responses to critical high-risk scenarios, or conversely, the misdirection of resources to low-risk conversations erroneously identified as high-risk. Additionally, users whose conversations are mistakenly flagged as high-risk could experience undue stress or anxiety, which could also lead to undeserved reputational harm, especially if actions are unjustly taken based on these incorrect classifications resulting from false positives (141; 381).

A key aspect for a safer real-world implementation of these algorithms could be through incorporating a human-agent reinforcement learning approach (239; 192; 15), where human insights and decision-making are integrated with the RL algorithm prioritization decisions, especially in cases where they believe a conversation has been misclassified. Through this approach, developers could use this feedback loop as an ongoing learning process for reinforcement learning and the conversation priority classification models from real-world interactions, which could significantly improve its accuracy over time. In addition, leveraging their feedback and communicating with them about how their feedback is being used to improve the algorithms could build trust and encourage more user engagement, as highlighted in prior research (144). More importantly, their involvement can facilitate the building of a more dynamic prioritizing system that could adapt and evolve based on real-world feedback and changing scenarios, rather than being static and potentially becoming outdated.

At the same time, this human-in-the-loop approach can encounter the inherent challenge of balancing timely intervention with thorough review. In situations requiring immediate interventions, the integration of human feedback with the RL algorithm could introduce delays. Therefore, we suggest the need for differentiating the timing and level of human involvement based on varying levels of priority. For instance, the high-priority cases that demand immediate action, automated systems can first provide rapid initial responses, potentially followed by more nuanced human evaluation and intervention, while for the lower-priority conversations, a more in-depth human review could be more appropriate to ensure it does not impede the real-time nature of the platform. As such, the real-world implementation of the human-in-the-loop approach needs careful structuring to maintain the essential real-time nature of communication platforms. A possible way to achieve this structuring could be through developing tiered response mechanisms where the level of human involvement is scaled according to the urgency and risk level of the conversation, similar to Meta's (Facebook) three-tiered system for managing its dangerous organizations and individuals policy (99).

Another critical component for ensuring a safer real-world implementation of the conversation prioritization approach could be the involvement of experts in auditing algorithms and overseeing the prioritization decisions (303). Auditing this prioritization approach would provide valuable insights into its real-world impacts on users and society by identifying and rectifying biases, inaccuracies, or unintended consequences that may not be apparent to the algorithm or its developers. In addition, regular auditing by experts would ensure that the prioritization decisions remain aligned with ethical standards and societal norms, thereby maintaining trust among youth and stakeholders. Thus, this audition would add a layer of accountability and transparency, which is essential in sensitive areas such as youth online safety and content moderation, where decisions could have significant implications for individuals and communities.

5.5.3 Limitations and Future Work

While our study offers significant insights into identifying conversations with high-risk priorities, its focus exclusively on Instagram poses a challenge to its wider applicability. To better understand the broader relevance and effectiveness of our approach, future research is encouraged to investigate other popular social media platforms among youth, like TikTok. In this study, we focused primarily on textual features to demonstrate the efficacy of our models, avoiding more complex features. Our future work will include contextual factors such as age, gender, and the nature of relationships, which have been shown to influence the performance of sexual content classifiers (286). Although we employed pre-trained risk detection models in this study, we plan to explore integrated multi-risk detection within this framework to assess how context identified by the RL agent affects risk detection accuracy and responsiveness in future work. Additionally, we plan to engage youth in evaluating the quality of our algorithmic classifiers by implementing a web-based risk detection system. This step will close the Human-Centered Machine Learning (HCML) loop, allowing direct user feedback on our algorithm's performance, essential for further refinement and real-world application and impact.

5.6 Conclusion

In today's digitally-driven world, youth are confronted with various online risks that affect both themselves and society. The necessity for timely and accurate risk identification is key for effective intervention by stakeholders like governments and online platforms. Our research marks a significant step in improving real-time responses, introducing a framework that emphasizes prioritizing high-risk situations. This innovative framework balances the urgency of addressing high-risk scenarios with resource constraints, offering a more focused approach to online risk management. This is particularly crucial in platforms with large volumes of interactions, where traditional automatic risk detection and moderation methods may struggle to keep pace. By prioritizing and continuously monitoring high-risk conversations, our framework could not only enhance the effectiveness of risk detection and intervention strategies but also significantly improve the overall safety of the younger generation.

CHAPTER 6

MOSafely, *Is that Sus?* A Youth-Centric Online Risk Assessment Dashboard

Citation: Ashwaq Alsoubai, Xavier Caddle, Ryan Doherty, Alexandra Koehler, Estefania Sanchez, Munmun De Choudhury, Pamela Wisniewski. MOSafely, Is that Sus? A Youth-Centric Online Risk Assessment Dashboard. Proceedings of the ACM on Computer-Supported Cooperative Work And Social Computing Demonstration (CSCW '22).

Current youth online safety and risk detection solutions are mostly geared toward parental control. As HCI researchers, we acknowledge the importance of leveraging a youth-centered approach when building Artificial Intelligence (AI) tools for adolescents' online safety. Therefore, we built the MOSafely, *Is that 'Sus' (youth slang for suspicious)?* a web-based risk detection assessment dashboard for youth (ages 13-21) to assess the AI risks identified within their online interactions (Instagram and Twitter Private conversations). This demonstration will showcase our novel system that embedded risk detection algorithms for youth evaluations and adopted the human–in–the loop approach for using youth evaluations to enhance the quality of machine learning models.

6.1 Introduction

Adolescents' social growth and developmental exploration are mainly mediated through extensive social media usage (21). Although social media provides youth with a unique opportunity to communicate and learn, it also exposes them to a wide array of risks that could have adverse consequences (207). A major trend of the current approaches for adolescents' online safety is relying on parental control that is not only privacy-invasive to youth but also overloads parents with unnecessary information (11; 131). Today, Artificial Intelligence (AI)-based risk detection technologies present promising potentials to automatically detect risky content (365). However, these models could pose a digital inequity especially for socio-economically disadvantaged youth (289). Thus, human-computer interaction (HCI) researchers have been advocating for building AI online safety solutions that are youth-centric (289). To address this, under the auspices of an initiative called Modus Operandi Safely (i.e., MOSafely), we built a youth-centered, web-based risk detection dashboard called, MOSafely,"*Is that Sus?,"* which leverages machine learning algorithms that we developed to detect risks within youth online interactions and provide them the ability to give feedback on the AI *suspected* risks.

6.2 Gaps in Existing Risk Detection Systems for Youth

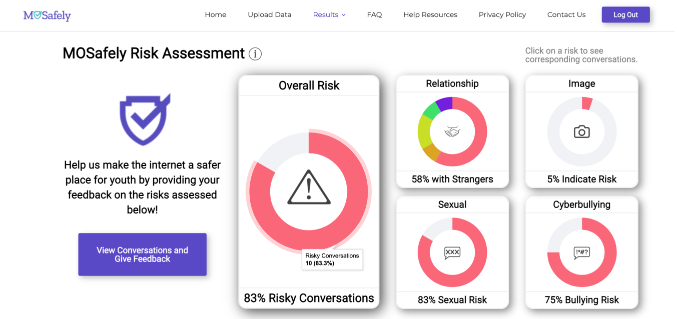
Most of the existing commercialized automatic risk detection solutions for youth have been social media platform-based that are not available for public use or evaluation (165). These solutions have also been mainly developed in isolation from youth's own perspective, resulting in high rates of false positives and hampering the potential of applying these algorithms in real life settings (301). Furthermore, the majority of the presented risk detection approaches in youth online safety literature lack youth engagement in these approaches (290; 179). The youths' perspective is important to be incorporated not only in identifying ground truth for the detection models, but also in enhancing the models' predictions based on their evaluations (290). In fact, recent research in Computer-Supported Cooperative Work and Social Computing (CSCW) has noted that human-centered approach in computing should leverage the personal, social, and cultural perspectives when designing and creating technological solutions (66). Therefore, the overarching aim of MOSafely's *Is that Sus?* dashboard is to address these limitations by applying a youth-centric approach to give youth more agency in their own online safety.

At the demo at CSCW, visitors will be able to navigate through one of the first novel initiatives to engage youth in the process of building AI risk detection systems. The presenter will have an opportunity to upload a sample Instagram and Twitter data files for the visitors to explore the features we designed for youths' evaluations and the machine learning algorithms we developed for multiple risks types (e.g., sexual messages and cyberbullying).

6.3 MOSafely, *Is that Sus?* Design Overview

MOSafely, *Is that Sus?* recognizes the importance of engaging youth when building online safety tools for them. It was designed to be customizable to allow teens upload social media files from different platforms such as Instagram and Twitter to address the current platform-based risk detection tools. We provide youth with step by step instructions about how to download their data from online platforms and then upload it our system. The following list describes the novelty of MOSafely, *Is that Sus?* risk detection assessment dashboard and how youths were engaged in assessing the AI risk predictions of their online conversations.

- **Theoretically grounded risk types.** To design MOSafely, *Is that Sus?*, we focused on the most prevalent risks that youth encounter online, which were sexual solicitation and cyberbullying (362). As such, we used systematic reviews of automatic (machine learning based) detection approaches for these risks (290; 179) and, accordingly designed the embedded algorithms. The algorithms classify a conversation as risky when sexual messages/solicitation and/or cyberbullying (text and image) were detected and safe when none of these risks were detected. Due to the importance of contextualizing the risks youth encounter to avoid unintentional harms (16), the relationship type classifier was designed to only

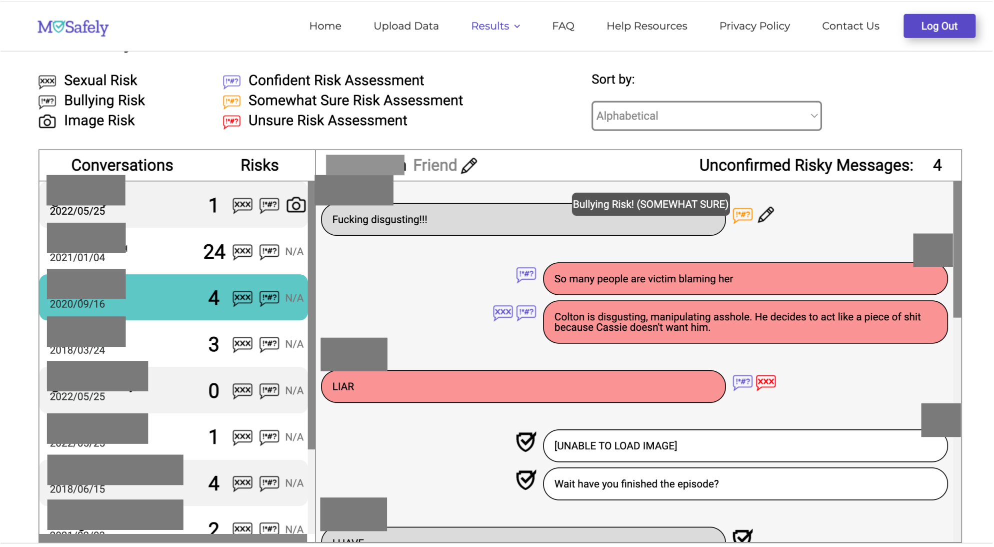


MOSafely Risk Assessment ⓘ

Click on a risk to see corresponding conversations.

Help us make the internet a safer place for youth by providing your feedback on the risks assessed below!

Figure 6.1: Screenshot of the risk assessment dashboard after users successfully uploaded their social media file.



XXX Sexual Risk
!*#? Bullying Risk
📷 Image Risk

!*#? Confident Risk Assessment
!*#? Somewhat Sure Risk Assessment
!*#? Unsure Risk Assessment

Sort by: Alphabetical

Conversations	Risks			
[redacted] 2022/05/25	1	XXX	!*#?	📷
[redacted] 2021/01/04	24	XXX	!*#?	N/A
[redacted] 2020/09/16	4	XXX	!*#?	N/A
[redacted] 2018/03/24	3	XXX	!*#?	N/A
[redacted] 2022/05/25	0	XXX	!*#?	N/A
[redacted]	1	XXX	!*#?	N/A
[redacted] 2018/06/15	4	XXX	!*#?	N/A
[redacted]	2	XXX	!*#?	N/A

[redacted] Friend ✎ Unconfirmed Risky Messages: 4

[redacted]
Fucking disgusting!!! — Bullying Risk! (SOMEWHAT SURE) !*#? ✎

[redacted]
!*#? So many people are victim blaming her
XXX !*#? Colton is disgusting, manipulating asshole. He decides to act like a piece of shit because Cassie doesn't want him.

[redacted]
LIAR !*#? XXX

[redacted]
[UNABLE TO LOAD IMAGE]
Wait have you finished the episode?

[redacted]

Figure 6.2: Screenshot of the conversation page, showing the edit icon for the message level feedback.

predict the relationship types (i.e., stranger, acquaintance, friend, significant other, family) without labeling whether the conversation is risky.

- **Embedded algorithms for user evaluation.** Due to the lack of existing solutions that embed machine learning algorithm for evaluation, we designed the MOSafely risk detection dashboard to be one of the first systems that embedded algorithms to be publicly available for youth evaluation. We developed our own machine learning algorithms to detect risks using the conversations and single messages to address a limitation in the current approaches that heavily rely on the conversation level as an input (290). Then these trained classifiers were integrated in MOSafely system to predict risks within the youth uploaded online interactions.

- **Teen-centric design to raise awareness.** MOSafely was designed for youth to review their online interactions to be more self-aware of the risky interactions they are having online. Therefore, the risk assessment dashboard was designed for them to have an at-a-glance overview of the AI detected risky conversations and navigate through them to reflect about what they found risky. The dashboard cards were designed to show the overall number of risky conversations as well as the number of conversations identified for different types of risks including sexual risks, cyberbullying, and relationship types. These cards are also useful for youth to filter their AI predicted risky conversations based on a risk type they found interesting or surprising.

- **Feedback mechanism to improve algorithms.** We leveraged a human-in-the-loop approach (239) to get feedback from end users (youth) on the accuracy of the risk predictions produced, which will be used by the system to further improve the accuracy of our trained algorithms. To this end, we designed a conversation page for the users to thoroughly review their conversations and give feedback on the AI detected risks. The conversation page allowed youth to submit feedback for predictions at the conversation level as well as the message level. Each conversation and message provided an overview of the risks, with a pop-up for feedback and contextual information (e.g., relationship type in conversations). Youth also had the option to provide written feedback with more details about why they disagreed with the predictions. The system also helped the youth keep track of their progress, by updating a "counter" which showed the number of risk assessment predictions not reviewed by the user yet. Ultimately, this feedback will enable us to compare the performances of conversations vs. message level algorithms for risk detection.

6.4 Technical Implementation

The following sections describe the development of machine learning algorithms and the AWS technical implementation.

6.4.1 Predictive Machine Learning Application Programming Interface

Due to the lack of publicly available pre-trained risk detection models, we developed and trained models to detect risks in youth online interactions. Prior to MOSafely risk detection dashboard, we collected an ecologically valid dataset that consisted of youth private conversations along with their risk flags to their conversations; (287) describe the design considerations behind this data collection. Starting from our other work that provided skeleton machine learning algorithms for risk detection, such as sexual risk and cyberbullying (13; 285; 178), we trained models using this dataset to detect risks including cyberbullying, sexual solicitation, and risky images for both conversation and message levels. For choosing the most accurate predictive models for each risk type, we trained traditional and deep learning models, with the best models were listed in Table 6.1. The best performing models for each risk type were then compiled as TensorFlow saved

Classification Level	**Classification Type**	**Model Type**	**Accuracy**	**F1**
Message	Sexual	DNN	87%	87%
	Cyberbullying		82%	82%
	Image	CNN	60%	89%
Conversation	Sexual	CNN	89%	90%
	Cyberbullying	LSTM	68%	63%
	Relationship Type	CNN	80%	89%

Table 6.1: Conversation and message level classifiers' accuracy results. CNN denotes Convolutional Neural Networks, DNN denotes Deep Neural Network, and LSTM denotes Long Short-Term Memory

models.

The modularized models were then hosted on the machine learning server (MLAPI). Since these are modularized, no retraining is needed each time the server runs the models. The main goal of the MLAPI server is to serve as an Application Programming Interface (API) that is scalable enough to incorporate several risk classifiers, to produce predictions for any text such as messages from phone message apps, and to be embedded in any online platform and/or mobile application to help youth navigate their own risks instantly. The MLAPI server responds to prediction requests with a JSON structured object containing fields which signify if the conversation is risky or non-risky. The response also includes the same fields *for each* distinct message in the conversation thereby providing conversation and message level risk prediction assessments.

6.4.2 AWS Backend

We used AWS Elastic Compute Cloud (EC2) to host the website that control the information flow between the web-front (users input) and the PHP back-end (data transmission to Database or storing social media folders in AWS S3 buckets). The MLAPI server used to store the trained machine learning models is hosted using an EC2 instance. The AWS Simple Storage Service (S3) was used to store the users' social media files. AWS Lambda function code was created and extended to parse the content of different social media platforms files.

The parsing process included converting the data file format (JSON or java script) to text and it also included sending the parsed conversations to the MLAPI to get the predictions. AWS Relational Database Service (RDS), a Health Insurance Portability and Accountability Act (HIPAA) (252) compliant service[1], was used to securely save users' social conversations and resulting risk assessment predictions in a password protected storage. The environment variables in Lambda functions and database passwords were encrypted using AWS Key Management Service to achieve at-rest and in-transit encryption. The RDS and Lambda

[1] https://docs.aws.amazon.com/whitepapers/latest/architecting-hipaa-security-and-compliance-on-aws/amazon-rds-for-sql-server.html

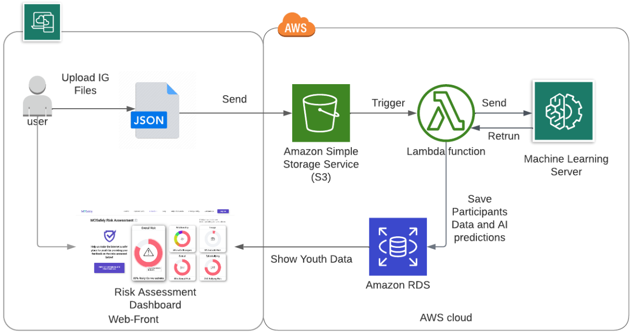


Figure 6.3: MOSafely Architecture.

functions were hosted under a Virtual Private Cloud (VPC) to protect the data transmission.

6.5 Future Research and Conclusion

MOSafely, *Is that Sus?* has not been formally evaluated by youth. Youths' feedback will be valuable to inform future research about the efficiency and applicability of the algorithms, as well as the intuitiveness of the presentation of the risk assessment predictions. We plan to perform a usability evaluation of this system with a subset of the youth population to resolve any design issues based on their workflow/usability standpoint (320). We also intend to investigate the perceived utility of the risk detection dashboard based on the perspectives of other stakeholders in youth online safety such as parents and youth social service providers.

While existing AI tools for youth online safety are mainly designed and developed behind corporate walls, we showcased MOSafely, *Is that Sus?* as a novel system that will open machine learning algorithms for public evaluation, especially from the youth. Youths' feedback and insights about the models' performances will be helpful in bringing to the market not only state-of-the-art but also youth-approved solutions for detecting risks they encounter online.

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.

7.1 Research Summary

In this dissertation study, we focused on providing insights on challenging and risky online interactions in which youth are dealing with, and used an ecologically valid dataset based on adolescents' perspectives on those types of risk and built classifiers to detect these risks in real-time. 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 real-time online 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 systematic literature review of 53 peer-reviewed articles on real-time risk detection on social media. Real-time detection was mainly operationalized as "early" detection after-the-fact based on pre-defined chunks of data and evaluated based on standard performance metrics, such as timeliness. We identified several human-centered opportunities for advancing current algorithms, such as integrating human insight in feature selection, algorithms' improvement considering human behavior, and utilizing human evaluations. This work serves as a critical call-to-action for the HCI and ML communities to work together to protect social media users before, during, and after exposure to risks.

- **RQ2- Study 1:***What insights can we gain regarding the online risk experiences of youth through their disclosures of sexual risk experiences when seeking peer support online?*

  In Chapter 3, e analyzed posts (N = 45, 955) made by adolescents (ages 13–17) on an online peer support platform to deeply examine their online sexual risk experiences. By applying a mixed methods approach, we 1) accurately (average of AUC = 0.90) identified posts that contained teen disclosures about online sexual risk experiences and classified the posts based on level of consent (i.e., consensual, non-consensual, sexual abuse) and relationship type (i.e., stranger, dating/friend, family) between the teen and the person in which they shared the sexual experience, 2) detected statistically significant differences in the proportions of posts based on these dimensions, and 3) further unpacked the nuance in how these online sexual risk experiences were typically characterized in the posts. Teens were significantly more likely to engage in consensual sexting with friends/dating partners; unwanted solicitations were more likely from strangers and sexual abuse was more likely when a family member was involved.

- **RQ3- Study 2:** *How does creating profiles of youth based on their self-reported online and offline risk behaviors inform about the multidimensionality of their lived risk experiences on social media?*
  In Chapter 4, We conducted a study with 173 adolescents (ages 13-21), who self-reported their offline and online risk ex- periences and uploaded their Instagram data to our study website to flag private conversations as unsafe. Risk profiles were first created based on the survey data and then compared with the risk-flagged social media data. Five risk profiles emerged: Low Risks (51% of the participants), Medium Risks (29%), Increased Sexting (8%), Increased Self-Harm (8%), and High Risks Perpetration (4%). Overall, the profiles correlated well with the social media data with the highest level of the risk occurring in the three smallest profiles. Youth who experienced increased sexting and self-harm frequently reported engaging in unsafe sexual conversations. Meanwhile, high risks perpetration was characterized by increased violence, threats, and sales/promotion of illegal activities. A key insight from our study was that offline risk behavior sometimes manifested differently in online contexts (i.e., offline self-harm as risky online sexual interactions). Our findings highlight the need for targeted risk prevention strategies for youth online safety

- **RQ4- Study 3:** *Can the insights gained from the prior two studies be built upon to create algorithms that accurately detect high-risk conversations in real-time?*
  In Chapter 6, we presented a novel two-level framework that first uses reinforcement learning to identify conversation stop points to prioritize messages for evaluation. Then, we optimize state-of-the- art deep learning models to accurately categorize risk priority (low, high). We apply this framework to a time-based simulation using a rich dataset of 23K private conversations with over 7 million messages do- nated by 194 youth (ages 13-21). We conducted an experiment comparing our new approach to a traditional conversation-level baseline. We found that the timeliness of conversations significantly improved from over 2 hours to approximately 16 minutes with only a slight reduction in accuracy (0.88 to 0.84). This study advances real-time detection approaches for social media data and provides a benchmark for future training reinforcement learning that prioritizes the timeliness of classifying high-risk conversations.

7.2 Research Contributions

Our research 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) as follows:

- 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 an in-depth synthesis of the current state-of-the-art, trends, and gaps in computational approaches for social media real-time risk detection and recommendations for a research agenda that would advance beyond the state-of-the-art within this research domain, which also is considered a survey contribution. Second, we provide a deeper understanding of sexual risk experiences for adolescents and uncover the nuances of these experiences that provided recommendations for future research and design.
- Study 2 makes a methodological contribution to investigate the multidimensional nature of youth online and offline risk experiences and created unique profiles of youth to uncover their salient risk experiences in their Instagram private conversations. These insights established the validity of youth risk annotations and make recommendations for future research and machine learning development.
- Our study 3 makes an artifact and methodological contributions by utilizing a human-centered approach by presenting a novel framework that utilizes Reinforcement Learning to dynamically identify when a conversation needs to be evaluated for high-risk priority.

7.3 Future Research Directions

For each study, the future research directions are stated in its corresponding chapter. In Chapter 6 of my dissertation, we introduced an innovative artificial interface designed to solicit feedback from young individuals on the efficacy of risk detection algorithms. Looking ahead, a pivotal avenue for future research lies in executing comprehensive user studies. These studies will be crucial not only for assessing the efficiency of the models in identifying risks but also for gauging youth engagement with these technologies. Specifically, a significant focus will be placed on evaluating the effectiveness of the study three framework in the context of high-risk conversations. Upon my return to my home country, I plan to extend this research by collecting and analyzing social media data from Saudi youth. This endeavor aims to deepen our understanding of the risk experiences unique to this demographic and to develop tailored risk detection solutions that resonate with the needs and nuances of Saudi Arabia's conservative culture. This bifocal approach will not only enhance our comprehension of global youth perspectives on digital safety but also pave the way for creating more culturally sensitive and effective risk detection mechanisms.

7.4 Conclusion

Overall, this work contributed to the dire societal issues of adolescent online safety by detecting and providing insights relevant to youths' risky online interactions and how to detect them in real time. 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 real-time risk detection to set a research agenda for the field. Additionally, it provides a deeper understanding of the multidimensionality in the youth population, which makes recommendations for designing targeted interventions for them and provides a novel framework that would pave the way for future research to enhance the real-time risk detection approaches for youth.

References

[1] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., et al. (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*.

[2] Abuse, S. and (SAMHSA), M. H. S. A. (2023). Samhsa announces national survey on drug use and health (nsduh) results detailing mental illness and substance use levels in 2021.

[3] Agardh, A., Cantor-Graae, E., and Östergren, P.-O. (2012). Youth, Sexual Risk-Taking Behavior, and Mental Health: a Study of University Students in Uganda. *Int'l Journal of Behavioral Medicine*, 19(2):208–216.

[4] Agha, Z., Badillo-Urquiola, K., and Wisniewski, P. J. (2023). " strike at the root": Co-designing real-time social media interventions for adolescent online risk prevention. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):1–32.

[5] Agha, Z., Ghaiumy Anaraky, R., Badillo-Urquiola, K., McHugh, B., and Wisniewski, P. (2021). '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.

[6] Agha, Z., Zhang, Z., Obajemu, O., Shirley, L., and J. Wisniewski, P. (2022). A case study on user experience bootcamps with teens to co-design real-time online safety interventions. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, pages 1–8.

[7] Agnew, R. (1992). Foundation for a general strain theory of crime and delinquency. *Criminology*, 30(1):47–88.

[8] Ahern, N. R., Kemppainen, J., and Thacker, P. (2016). Awareness and knowledge of child and adolescent risky behaviors: A parent's perspective. *Journal of Child and Adolescent Psychiatric Nursing*, 29(1):6–14.

[9] Ahmed, Y. A., Ahmad, M. N., Ahmad, N., and Zakaria, N. H. (2019). Social media for knowledge-sharing: A systematic literature review. *Telematics and informatics*, 37:72–112.

[10] Ajmani, L. H., Chancellor, S., Mehta, B., Fiesler, C., Zimmer, M., and De Choudhury, M. (2023). A systematic review of ethics disclosures in predictive mental health research. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1311–1323.

[11] Akter, M., Godfrey, A. J., Kropczynski, J., Lipford, H. R., and Wisniewski, P. J. (2022). From parental control to joint family oversight: Can parents and teens manage mobile online safety and privacy as equals? *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW1):1–28.

[12] Al-Garadi, M. A., Hussain, M. R., Khan, N., Murtaza, G., Nweke, H. F., Ali, I., Mujtaba, G., Chiroma, H., Khattak, H. A., and Gani, A. (2019). Predicting cyberbullying on social media in the big data era using machine learning algorithms: review of literature and open challenges. *IEEE Access*, 7:70701–70718.

[13] Ali, S., Razi, A., Kim, S., Alsoubai, A., Gracie, J., De Choudhury, M., Wisniewski, P. J., and Stringhini, G. (2022). Understanding the digital lives of youth: Analyzing media shared within safe versus unsafe private conversations on instagram. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–14.

[14] Ali, S., Razi, A., Kim, S., Alsoubai, A., Ling, C., De Choudhury, M., Wisniewski, P. J., and Stringhini, G. (2023). Getting meta: A multimodal approach for detecting unsafe conversations within instagram direct messages of youth. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):1–30.

[15] Alsoubai, A., Caddle, X. V., Doherty, R., Koehler, A. T., Sanchez, E., De Choudhury, M., and Wisniewski, P. J. (2022). Mosafely, is that sus? a youth-centric online risk assessment dashboard. In *Companion Publication of the 2022 Conference on Computer Supported Cooperative Work and Social Computing*, pages 197–200.

[16] AlSoubai, A., Song, J., Razi, A., Naher, N., De Choudhury, M., and Wisniewski, P. J. (2022). From 'friends with benefits' to 'sextortion:' a nuanced investigation of adolescents' online sexual Risk experiences.

[17] Alsoubai, A., Song, J., Razi, A., Naher, N., De Choudhury, M., and Wisniewski, P. J. (2022). From'friends with benefits' to'sextortion:'a nuanced investigation of adolescents' online sexual risk experiences. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–32.

[18] Alsoubai, Ashwaq; RAZI, A., AGHA, Z., ALI, S., STRINGHINI, G., DE CHODHURY, M., and WISNIEWSKI, P. J. (2024). Profiling the offline and online risk experiences of youth to develop targeted interventions for online safety.

[19] Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., et al. (2019). Guidelines for human-ai interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pages 1–13.

[20] Andalibi, N., Haimson, O. L., De Choudhury, M., and Forte, A. (2016). Understanding social media disclosures of sexual abuse through the lenses of support seeking and anonymity. In *Proceedings of the 2016 CHI Conf. on human factors in computing systems*, pages 3906–3918, San Jose, CA, USA. ACM.

[21] Anderson, M. and Jiang, J. (2018). Teens, Social Media & Technology 2018 | Pew Research Center.

[22] Ansary, N. S. (2020). Cyberbullying: Concepts, theories, and correlates informing evidence-based best practices for prevention. *Aggression and violent behavior*, 50:101343.

[23] Aragon, C., Guha, S., Kogan, M., Muller, M., and Neff, G. (2022). *Human-centered data science: an introduction*. MIT Press.

[24] Arango, C., Díaz-Caneja, C. M., McGorry, P. D., Rapoport, J., Sommer, I. E., Vorstman, J. A., McDaid, D., Marín, O., Serrano-Drozdowskyj, E., Freedman, R., et al. (2018). Preventive strategies for mental health. *The Lancet Psychiatry*, 5(7):591–604.

[25] Arendt, F., Scherr, S., and Romer, D. (2019). Effects of exposure to self-harm on social media: Evidence from a two-wave panel study among young adults. *New Media & Society*, 21(11-12):2422–2442.

[26] Arif, M. (2021). A systematic review of machine learning algorithms in cyberbullying detection: future directions and challenges. *Journal of Information Security and Cybercrimes Research*, 4(1):01–26.

[27] Atske, S. (2022). Teens and cyberbullying 2022.

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

[29] Ayinmoro, A. D., Uzobo, E., Teibowei, B. J., and Fred, J. B. (2020). Sexting and other risky sexual behaviour among female students in a nigerian academic institution. *Journal of Taibah University Medical Sciences*, 15(2):116–121.

[30] Backe, E. L., Lilleston, P., and McCleary-Sills, J. (2018). Networked individuals, gendered violence: A literature review of cyberviolence. *Violence and gender*, 5(3):135–146.

[31] Bailey, S., Camlin, C., and Ennett, S. (1998). Substance use and risky sexual behavior among homeless and runaway youth. *Journal of Adolescent Health*, 23(6):378–388.

[32] Barnhart, B. (2022). 41 of the most important social media marketing statistics for 2022.

[33] Bassen, J., Balaji, B., Schaarschmidt, M., Thille, C., Painter, J., Zimmaro, D., Games, A., Fast, E., and Mitchell, J. C. (2020). Reinforcement learning for the adaptive scheduling of educational activities. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–12.

[34] Baumer, E. P. (2017). Toward human-centered algorithm design. *Big Data & Society*, 4(2):2053951717718854.

[35] Baumer, E. P., Guha, S., Quan, E., Mimno, D., and Gay, G. K. (2015). Missing photos, suffering withdrawal, or finding freedom? how experiences of social media non-use influence the likelihood of reversion. *Social Media+ Society*, 1(2):2056305115614851.

[36] Baumgartner, S. E., Sumter, S. R., Peter, J., and Valkenburg, P. M. (2012). Identifying teens at risk: Developmental pathways of online and offline sexual risk behavior. *Pediatrics*, 130(6):e1489–e1496.

[37] Baumgartner, S. E., Valkenburg, P. M., and Peter, J. (2010). Unwanted online sexual solicitation and risky sexual online behavior across the lifespan. *Journal of Applied Developmental Psychology*, 31(6):439–447.

[38] Bello-Orgaz, G., Jung, J. J., and Camacho, D. (2016). Social big data: Recent achievements and new challenges. *Information Fusion*, 28:45–59.

[39] Benjamin, R. (2019). Assessing risk, automating racism. *Science*, 366(6464):421–422.

[40] Benotsch, E. G., Snipes, D. J., Martin, A. M., and Bull, S. S. (2013). Sexting, substance use, and sexual risk behavior in young adults. *Journal of adolescent health*, 52(3):307–313.

[41] Beres, M. A. (2007). 'spontaneous' sexual consent: An analysis of sexual consent literature. *Feminism & Psychology*, 17(1):93–108.

[42] Beres, M. A. (2014). Rethinking the concept of consent for anti-sexual violence activism and education. *Feminism & Psychology*, 24(3):373–389.

[43] Bhargava, R., Sharma, Y., and Sharma, S. (2016). Sentiment analysis for mixed script indic sentences. In *2016 Int'l Conf. on advances in computing, communications and informatics (ICACCI)*, pages 524–529. IEEE.

[44] Bianchi, D., Morelli, M., Nappa, M. R., Baiocco, R., and Chirumbolo, A. (2021). A bad romance: Sexting motivations and teen dating violence. *Journal of interpersonal violence*, 36(13-14):6029–6049.

[45] Bishop, A. S., Fleming, C. M., and Nurius, P. S. (2020). Substance use profiles among gang-involved youth: social ecology implications for service approaches. *Children and youth services review*, 119:105600.

[46] Black, P. J., Woodworth, M., Tremblay, M., and Carpenter, T. (2012). A review of trauma-informed treatment for adolescents. *Canadian Psychology/Psychologie canadienne*, 53(3):192–203. Place: US Publisher: Educational Publishing Foundation.

[47] Blei, D., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation: the journal of machine learning research, v. 3.

[48] Bours, P. and Kulsrud, H. (2019). Detection of cyber grooming in online conversation. In *2019 IEEE Int'l Workshop on Information Forensics and Security (WIFS)*, pages 1–6. IEEE.

[49] Boyle, S. C. and LaBrie, J. W. (2021). A gamified, social media–inspired, web-based personalized normative feedback alcohol intervention for lesbian, bisexual, and queer-identified women: Protocol for a hybrid trial. *JMIR Research Protocols*, 10(4):e24647.

[50] Braun, V. and Clarke, V. (2012). *Thematic analysis.* American Psychological Association.

[51] Breuner, C. C., Mattson, G., Adelman, W. P., Alderman, E. M., Garofalo, R., Marcell, A. V., Powers, M. E., Upadhya, K. K., Yogman, M. W., Bauer, N. S., et al. (2016). Sexuality education for children and adolescents. *Pediatrics*, 138(2).

[52] Brinkley, D. Y., Ackerman, R. A., Ehrenreich, S. E., and Underwood, M. K. (2017). Sending and receiving text messages with sexual content: Relations with early sexual activity and borderline personality features in late adolescence. *Computers in human behavior*, 70:119–130.

[53] Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. (2016). Openai gym. *arXiv preprint arXiv:1606.01540*.

[54] Brown, G. and Yule, G. (1983). *Discourse analysis*. Cambridge university press.

[55] Brown, R. C., Fischer, T., Goldwich, A. D., Keller, F., Young, R., and Plener, P. L. (2018). # cutting: Non-suicidal self-injury (nssi) on instagram. *Psychological medicine*, 48(2):337–346.

[56] Bruckman, A. (2002). 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.

[57] Bruda, S. D. and Akl, S. G. (2003). Real-time computation: A formal definition and its applications. *International Journal of Computers and Applications*, 25(4):247–257.

[58] Burdisso, S. G., Errecalde, M., and Montes-y Gómez, M. (2019). A text classification framework for simple and effective early depression detection over social media streams. *Expert Systems with Applications*, 133:182–197.

[59] Burén, J. and Lunde, C. (2018). Sexting among adolescents: A nuanced and gendered online challenge for young people. *Computers in Human Behavior*, 85:210–217.

[60] Caddle, X. V., Naher, N., Miller, Z. P., Badillo-Urquiola, K., and Wisniewski, P. J. (2023). Duty to respond: The challenges social service providers face when charged with keeping youth safe online. *Proceedings of the ACM on Human-Computer Interaction*, 7(GROUP):1–35.

[61] Caddle, X. V., Razi, A., Kim, S., Ali, S., Popo, T., Stringhini, G., De Choudhury, M., and Wisniewski, P. J. (2021). Mosafely: Building an open-source hcai community to make the internet a safer place for youth. In *Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing*, pages 315–318.

[62] Calvete, E., Fernández-González, L., Royuela-Colomer, E., Morea, A., Larrucea-Iruretagoyena, M., Machimbarrena, J., Gónzalez-Cabrera, J., and Orue, I. (2021). Moderating factors of the association between being sexually solicited by adults and active online sexual behaviors in adolescents. *Computers in Human Behavior*, page 106935.

[63] Cano, A. E., Fernandez, M., and Alani, H. (2014). Detecting child grooming behaviour patterns on social media. In *Social Informatics: 6th International Conference, SocInfo 2014, Barcelona, Spain, November 11-13, 2014. Proceedings 6*, pages 412–427. Springer.

[64] Carey, J. L., Carreiro, S., Chapman, B., Nader, N., Chai, P. R., Pagoto, S., and Jake-Schoffman, D. E. (2018). Some and self harm: The use of social media in depressed and suicidal youth. In *Proceedings of the... Annual Hawaii International Conference on System Sciences. Annual Hawaii International Conference on System Sciences*, volume 2018, page 3314. NIH Public Access.

[65] Chancellor, S. (2023). Toward practices for human-centered machine learning. *Communications of the ACM*, 66(3):78–85.

[66] Chancellor, S., Baumer, E. P., and De Choudhury, M. (2019a). Who is the" human" in human-centered machine learning: The case of predicting mental health from social media. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–32.

[67] Chancellor, S., Birnbaum, M. L., Caine, E. D., Silenzio, V. M., and De Choudhury, M. (2019b). A taxonomy of ethical tensions in inferring mental health states from social media. In *Proceedings of the Conf. on fairness, accountability, and transparency*, pages 79–88.

[68] Chang, J. P., Schluger, C., and Danescu-Niculescu-Mizil, C. (2022). Thread with caution: Proactively helping users assess and deescalate tension in their online discussions. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–37.

[69] Chatzakou, D., Leontiadis, I., Blackburn, J., Cristofaro, E. D., Stringhini, G., Vakali, A., and Kourtellis, N. (2019). Detecting cyberbullying and cyberaggression in social media. *ACM Transactions on the Web (TWEB)*, 13(3):1–51.

[70] Chaudhary, M., Saxena, C., and Meng, H. (2021). Countering online hate speech: An nlp perspective. *arXiv preprint arXiv:2109.02941*.

[71] Chaudhary, P., Peskin, M., Temple, J. R., Addy, R. C., Baumler, E., and Ross, S. (2017). Sexting and mental health: a school-based longitudinal study among youth in texas. *Journal of Applied Research on Children*, 8(1):11.

[72] Chelmis, C. and Zois, D.-S. (2021). Dynamic, incremental, and continuous detection of cyberbullying in online social media. *ACM Transactions on the Web (TWEB)*, 15(3):1–33.

[73] Chen, J., Mullins, C. D., Novak, P., and Thomas, S. B. (2016). Personalized strategies to activate and empower patients in health care and reduce health disparities. *Health Education & Behavior*, 43(1):25–34.

[74] Chen, J. X., McDonald, A., Zou, Y., Tseng, E., Roundy, K. A., Tamersoy, A., Schaub, F., Ristenpart, T., and Dell, N. (2022). Trauma-informed computing: Towards safer technology experiences for all. In *Proceedings of the 2022 CHI conference on human factors in computing systems*, pages 1–20.

[75] Chen, T., Li, X., Yin, H., and Zhang, J. (2018). Call attention to rumors: Deep attention based recurrent neural networks for early rumor detection. In *Trends and Applications in Knowledge Discovery and Data Mining: PAKDD 2018 Workshops, BDASC, BDM, ML4Cyber, PAISI, DaMEMO, Melbourne, VIC, Australia, June 3, 2018, Revised Selected Papers 22*, pages 40–52. Springer.

[76] Chen, X.-W. and Lin, X. (2014). Big data deep learning: challenges and perspectives. *IEEE access*, 2:514–525.

[77] Cheng, L., Guo, R., Silva, Y. N., Hall, D., and Liu, H. (2021). Modeling temporal patterns of cyberbullying detection with hierarchical attention networks. *ACM/IMS Transactions on Data Science*, 2(2):1–23.

[78] Cheng, L., Li, J., Silva, Y. N., Hall, D. L., and Liu, H. (2019). Xbully: Cyberbullying detection within a multi-modal context. In *Proceedings of the twelfth acm international conference on web search and data mining*, pages 339–347.

[79] Cho, E. and Kim, S. (2015). Cronbach's coefficient alpha: Well known but poorly understood. *Organizational research methods*, 18(2):207–230.

[80] Chouzenoux, E. and Pesquet, J.-C. (2017). A stochastic majorize-minimize subspace algorithm for online penalized least squares estimation. *IEEE Transactions on Signal Processing*, 65(18):4770–4783.

[81] Chowdhury, A. G., Sawhney, R., Mathur, P., Mahata, D., and Shah, R. R. (2019). 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.

[82] Citron, D. K. and Franks, M. A. (2014). Criminalizing revenge porn. *Wake Forest L. Rev.*, 49:345.

[83] Cohen, J. (1968). Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213.

[84] Collins, W. A. and Laursen, B. (2004). Changing relationships, changing youth: Interpersonal contexts of adolescent development. *The Journal of Early Adolescence*, 24(1):55–62.

[85] Congress (2023). S.1409 - kids online safety act 118th congress (2023-2024).

[86] Connolly, J. and Goldberg, A. (1999). Romantic relationships in adolescence: The role of friends and peers in their emergence and development.

[87] Crestani, F., Losada, D. E., and Parapar, J. (2022). *Early Detection of Mental Health Disorders by Social Media Monitoring: The First Five Years of the ERisk Project*, volume 1018. Springer Nature.

[88] Cromer, L. D. and Newman, E. (2011). Research ethics in victimization studies: Widening the lens. *Violence against women*, 17(12):1536–1548.

[89] Cronbach, L. J. and Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological bulletin*, 52(4):281.

[90] Cuevas, A., Febrero, M., and Fraiman, R. (2004). An anova test for functional data. *Computational statistics & data analysis*, 47(1):111–122.

[91] Cui, Y., Jia, M., Lin, T.-Y., Song, Y., and Belongie, S. (2019). Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9268–9277.

[92] Dahiya, S., Sharma, S., Sahnan, D., Goel, V., Chouzenoux, E., Elvira, V., Majumdar, A., Bandhakavi, A., and Chakraborty, T. (2021). Would your tweet invoke hate on the fly? forecasting hate intensity of reply threads on twitter. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 2732–2742.

[93] Das, A., Liu, H., Kovatchev, V., and Lease, M. (2023). The state of human-centered nlp technology for fact-checking. *Information Processing Management*, 60(2):103219.

[94] David, A. (2017). Self-reflection in illness and health: literal and metaphorical? *Palgrave Communications*, 3(1):1–6.

[95] David-Ferdon, C., Vivolo-Kantor, A. M., Dahlberg, L. L., Marshall, K. J., Rainford, N., and Hall, J. E. (2016). A comprehensive technical package for the prevention of youth violence and associated risk behaviors.

[96] De Santisteban, P. and Gámez-Guadix, M. (2018). Prevalence and risk factors among minors for online sexual solicitations and interactions with adults. *The Journal of Sex Research*, 55(7):939–950.

[97] Del Rey, R., Ojeda, M., Casas, J. A., Mora-Merchán, J. A., and Elipe, P. (2019). Sexting among adolescents: the emotional impact and influence of the need for popularity. *Frontiers in psychology*, 10:1828.

[98] Dhanasekaran, P., Srinivasan, H., Sree, S. S., Devi, I. S. G., Sankar, S., and Vijayaraghavan, V. (2021). Somps-net: Attention based social graph framework for early detection of fake health news. In *Data Mining: 19th Australasian Conference on Data Mining, AusDM 2021, Brisbane, QLD, Australia, December 14-15, 2021, Proceedings*, pages 165–179. Springer.

[99] Díaz, Á. and Hecht-Felella, L. (2021). Double standards in social media content moderation. *Brennan Center for Justice at New York University School of Law. https://www. brennancenter. org/our-work/research-reports/double-standards-socialmedia-content-moderation*.

[100] Dir, A. L., Cyders, M. A., and Coskunpinar, A. (2013). From the bar to the bed via mobile phone: A first test of the role of problematic alcohol use, sexting, and impulsivity-related traits in sexual hookups. *Computers in Human Behavior*, 29(4):1664–1670.

[101] Dodhiawala, R. T., Sridharan, N., Raulefs, P., and Pickering, C. (1989). Real-time ai systems: A definition and an architecture. In *IJCAI*, pages 256–264. Citeseer.

[102] Döring, N. (2014). Consensual sexting among adolescents: Risk prevention through abstinence education or safer sexting? *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, 8(1).

[103] Doshi-Velez, F. and Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

[104] Dowlagar, S. and Mamidi, R. (2021). A survey of recent neural network models on code-mixed indian hate speech data. In *Forum for Information Retrieval Evaluation*, pages 67–74.

[105] Doyle, C., Douglas, E., and O'Reilly, G. (2021). The outcomes of sexting for children and adolescents: A systematic review of the literature. *Journal of adolescence*, 92:86–113.

[106] Dredge, R. and Schreurs, L. (2020). Social media use and offline interpersonal outcomes during youth: A systematic literature review. *Mass Communication and Society*, 23(6):885–911.

[107] Duan, S., Duan, Z., Li, R., Wilson, A., Wang, Y., Jia, Q., Yang, Y., Xia, M., Wang, G., Jin, T., et al. (2020). Bullying victimization, bullying witnessing, bullying perpetration and suicide risk among adolescents: A serial mediation analysis. *Journal of affective disorders*, 273:274–279.

[108] Dym, B., Brubaker, J. R., Fiesler, C., and Semaan, B. (2019). "coming out okay": Community narratives for lgbtq identity recovery work. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–28.

[109] Ebrahimi, M., Suen, C. Y., and Ormandjieva, O. (2016). Detecting predatory conversations in social media by deep convolutional neural networks. *Digital Investigation*, 18:33–49.

[110] Eiband, M., Schneider, H., Bilandzic, M., Fazekas-Con, J., Haug, M., and Hussmann, H. (2018). Bringing transparency design into practice. In *23rd international conference on intelligent user interfaces*, pages 211–223.

[111] Eisenstein, J., Ahmed, A., and Xing, E. P. (2011). Sparse additive generative models of text. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 1041–1048.

[112] Elliott, V., Christopher, N., Deck, A., and Schwartz, L. (2021). The facebook papers reveal staggering failures in the global south. rest of world.

[113] ElSherief, M., Kulkarni, V., Nguyen, D., Wang, W. Y., and Belding, E. (2018). Hate lingo: A target-based linguistic analysis of hate speech in social media. In *Proceedings of the international AAAI conference on web and social media*, volume 12.

[114] Escalante, H. J., Villatoro-Tello, E., Garza, S. E., López-Monroy, A. P., Montes-y Gómez, M., and Villaseñor-Pineda, L. (2017). Early detection of deception and aggressiveness using profile-based representations. *Expert Systems with Applications*, 89:99–111.

[115] Faber, B., Michelet, G. C., Weidmann, N., Mukkamala, R. R., and Vatrapu, R. (2019). Bpdims: A blockchain-based personal data and identity management system.

[116] Facebook (2023). How facebook ai helps suicide prevention.

[117] Fiesler, C., Brubaker, J. R., Forte, A., Guha, S., McDonald, N., and Muller, M. (2019). Qualitative methods for cscw: Challenges and opportunities. In *Conf. Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing*, pages 455–460.

[118] Finkelhor, D., Walsh, K., Jones, L., Mitchell, K., and Collier, A. (2021). Youth internet safety education: aligning programs with the evidence base. *Trauma, violence, & abuse*, 22(5):1233–1247.

[119] Foundation, T. N. (2022). Online safety (for teens) - nemours kidshealth.

[120] Founta, A. M., Chatzakou, D., Kourtellis, N., Blackburn, J., Vakali, A., and Leontiadis, I. (2019). A unified deep learning architecture for abuse detection. In *Proceedings of the 10th ACM conference on web science*, pages 105–114.

[121] Frankel, A. S., Bass, S. B., Patterson, F., Dai, T., and Brown, D. (2018). Sexting, risk behavior, and mental health in adolescents: An examination of 2015 pennsylvania youth risk behavior survey data. *Journal of school health*, 88(3):190–199.

[122] Fredlund, C., Svedin, C. G., Priebe, G., Jonsson, L., and Wadsby, M. (2017). Self-reported frequency of sex as self-injury (SASI) in a national study of Swedish adolescents and association to sociodemographic factors, sexual behaviors, abuse and mental health. *Child and Adolescent Psychiatry and Mental Health*, 11(1):9.

[123] Fung, I. C.-H., Blankenship, E. B., Ahweyevu, J. O., Cooper, L. K., Duke, C. H., Carswell, S. L., Jackson, A. M., Jenkins III, J. C., Duncan, E. A., Liang, H., et al. (2020). Public health implications of image-based social media: a systematic review of instagram, pinterest, tumblr, and flickr. *The Permanente Journal*, 24.

[124] Gámez-Guadix, M. and Mateos-Pérez, E. (2019). Longitudinal and reciprocal relationships between sexting, online sexual solicitations, and cyberbullying among minors. *Computers in Human Behavior*, 94:70–76.

[125] Ganesh, B. and Bright, J. (2020). Countering extremists on social media: Challenges for strategic communication and content moderation.

[126] Gardner, P. L. (1996). The dimensionality of attitude scales: a widely misunderstood idea. *International Journal of Science Education*, 18(8):913–919.

[127] Gassó, A. M., Klettke, B., Agustina, J. R., and Montiel, I. (2019). Sexting, Mental Health, and Victimization Among Adolescents: A Literature Review. *Int'l Journal of Environmental Research and Public Health*, 16(13):2364. Number: 13 Publisher: Multidisciplinary Digital Publishing Institute.

[128] (GDPR), G. D. P. R. (2021). Art. 20 gdpr – right to data portability — general data protection regulation (gdpr).

[129] Ge, S., Cheng, L., and Liu, H. (2021). Improving cyberbullying detection with user interaction. In *Proceedings of the Web Conference 2021*, pages 496–506.

[130] Gewirtz-Meydan, A., Mitchell, K. J., and Rothman, E. F. (2018). What do kids think about sexting? *Computers in Human Behavior*, 86:256–265.

[131] Ghosh, A. K., Badillo-Urquiola, K., Rosson, M. B., Xu, H., Carroll, J., and Wisniewski, P. J. (2018). A matter of control or safety? Examining parental use of technical monitoring apps on teens' mobile devices. In *CHI 2018 - Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems: Engage with CHI*. Association for Computing Machinery.

[132] Giumetti, G. W. and Kowalski, R. M. (2022). Cyberbullying via social media and well-being. *Current Opinion in Psychology*, page 101314.

[133] Golbeck, J., Ashktorab, Z., Banjo, R. O., Berlinger, A., Bhagwan, S., Buntain, C., Cheakalos, P., Geller, A. A., Gnanasekaran, R. K., Gunasekaran, R. R., et al. (2017). A large labeled corpus for online harassment research. In *Proceedings of the 2017 ACM on web science conference*, pages 229–233.

[134] Gong, C., Saha, K., and Chancellor, S. (2021). " the smartest decision for my future": Social media reveals challenges and stress during post-college life transition. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–29.

[135] Görzig, A. (2016). Adolescents' experience of offline and online risks: Separate and joint propensities. *Computers in Human Behavior*, 56:9–13.

[136] Gradinger, P., Strohmeier, D., and Spiel, C. (2009). Traditional bullying and cyberbullying: Identification of risk groups for adjustment problems. *Zeitschrift für Psychologie/Journal of Psychology*, 217(4):205–213.

[137] Grimmer, J. and Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis*, 21(3):267–297.

[138] Guha, S., Baumer, E. P., and Gay, G. K. (2018). Regrets, i've had a few: When regretful experiences do (and don't) compel users to leave facebook. In *proceedings of the 2018 ACM Conf. on Supporting Groupwork*, pages 166–177.

[139] Gunawan, F. E., Ashianti, L., Candra, S., and Soewito, B. (2016). Detecting online child grooming conversation. In *2016 11th Int'l Conf. on Knowledge, Information and Creativity Support Systems (KICSS)*, pages 1–6. IEEE.

[140] Guntuku, S. C., Yaden, D. B., Kern, M. L., Ungar, L. H., and Eichstaedt, J. C. (2017). Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences*, 18:43–49.

[141] Haimson, O. L., Delmonaco, D., Nie, P., and Wegner, A. (2021). Disproportionate removals and differing content moderation experiences for conservative, transgender, and black social media users: Marginalization and moderation gray areas. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–35.

[142] Hammer, M. A., Dunfield, J., Headley, K., Labich, N., Foster, J. S., Hicks, M., and Van Horn, D. (2015). Incremental computation with names. In *Proceedings of the 2015 ACM SIGPLAN International Conference on Object-Oriented Programming, Systems, Languages, and Applications*, pages 748–766.

[143] Hammer, M. A., Phang, K. Y., Hicks, M., and Foster, J. S. (2014). Adapton: Composable, demand-driven incremental computation. *ACM SIGPLAN Notices*, 49(6):156–166.

[144] Hamon, R., Junklewitz, H., Sanchez, I., Malgieri, G., and De Hert, P. (2022). Bridging the gap between ai and explainability in the gdpr: towards trustworthiness-by-design in automated decision-making. *IEEE Computational Intelligence Magazine*, 17(1):72–85.

[145] Hanna, A., Denton, E., Smart, A., and Smith-Loud, J. (2020). Towards a critical race methodology in algorithmic fairness. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 501–512.

[146] Hartikainen, H., Razi, A., and Wisniewski, P. (2021a). Safe sexting: The advice and support adolescents receive from peers regarding online sexual risks. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–31.

[147] Hartikainen, H., Razi, A., and Wisniewski, P. (2021b). 'if you care about me, you'll send me a pic'-examining the role of peer pressure in adolescent sexting. In *Companion Publication of the 2021 Conf. on Computer Supported Cooperative Work and Social Computing*, pages 67–71.

[148] Hartley, K. and Vu, M. K. (2020). Fighting fake news in the covid-19 era: policy insights from an equilibrium model. *Policy sciences*, 53(4):735–758.

[149] Hartson, R. and Pyla, P. S. (2018). *The UX book: Agile UX design for a quality user experience*. Morgan Kaufmann.

[150] Hassan, N., Poudel, A., Hale, J., Hubacek, C., Huq, K. T., Santu, S. K. K., and Ahmed, S. I. (2020). Towards automated sexual violence report tracking. In *Proceedings of the Int'l AAAI Conf. on Web and Social Media*, volume 14, pages 250–259.

[151] He, Y., Li, J., Song, Y., He, M., Peng, H., et al. (2018). Time-evolving text classification with deep neural networks. In *IJCAI*, volume 18, pages 2241–2247.

[152] Hébert, M., Amédée, L. M., Théorêt, V., and Petit, M.-P. (2021). Diversity of adaptation profiles in youth victims of child sexual abuse. *Psychological trauma: theory, research, practice, and policy*.

[153] Henderson, L. (2011). Sexting and sexual relationships among teens and young adults. *McNair Scholars Research Journal*, 7(1):9.

[154] Henry, N. and Powell, A. (2015). Embodied harms: Gender, shame, and technology-facilitated sexual violence. *Violence against women*, 21(6):758–779.

[155] Henry, N. and Powell, A. (2016). Sexual violence in the digital age: The scope and limits of criminal law. *Social & legal studies*, 25(4):397–418.

[156] Hildebrandt, M. (2008). Defining profiling: a new type of knowledge? In *Profiling the European citizen*, pages 17–45. Springer.

[157] Hiraman, B. R. et al. (2018). A study of apache kafka in big data stream processing. In *2018 International Conference on Information, Communication, Engineering and Technology (ICICET)*, pages 1–3. IEEE.

[158] Hsieh, H.-F. and Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative health research*, 15(9):1277–1288.

[159] Huang, H., Zhou, L., and Jiang, Y. (2021). Early detection of fake news based on multiple information features. In *2021 4th International Conference on Data Science and Information Technology*, pages 414–419.

[160] Hudson, H. K. and Fetro, J. V. (2015). Sextual activity: Predictors of sexting behaviors and intentions to sext among selected undergraduate students. *Computers in Human Behavior*, 49:615–622.

[161] Instagram (2021). Continuing to make instagram safer for the youngest members of our community.

[162] Isah, H., Abughofa, T., Mahfuz, S., Ajerla, D., Zulkernine, F., and Khan, S. (2019). A survey of distributed data stream processing frameworks. *IEEE Access*, 7:154300–154316.

[163] Jaimes, A., Gatica-Perez, D., Sebe, N., and Huang, T. S. (2007). Guest editors' introduction: Human-centered computing–toward a human revolution. *Computer*, 40(5):30–34.

[164] Jhaver, S., Birman, I., Gilbert, E., and Bruckman, A. (2019). Human-machine collaboration for content regulation: The case of reddit automoderator. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 26(5):1–35.

[165] Jia, H., Wisniewski, P. J., Xu, H., Rosson, M. B., and Carroll, J. M. (2015). 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*, pages 583–599.

[166] Jonzon, E. and Lindblad, F. (2004). Disclosure, reactions, and social support: Findings from a sample of adult victims of child sexual abuse. *Child maltreatment*, 9(2):190–200.

[167] Kaluarachchi, T. (2021). Human-centered machine learning.

[168] Kamburugamuve, S., Fox, G., Leake, D., and Qiu, J. (2013). Survey of distributed stream processing for large stream sources. *Grids Ucs Indiana Edu*, 2:1–16.

[169] Karasavva, V. (2020). *iPredator: Image-based Sexual Abuse Risk Factors and Motivators*. PhD thesis, Carleton University.

[170] Karlekar, S. and Bansal, M. (2018). Safecity: Understanding diverse forms of sexual harassment personal stories. *arXiv preprint arXiv:1809.04739*.

[171] Katell, M., Young, M., Dailey, D., Herman, B., Guetler, V., Tam, A., Bintz, C., Raz, D., and Krafft, P. (2020). Toward situated interventions for algorithmic equity: lessons from the field. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 45–55.

[172] Kern, R. (2022). Push to rein in social media sweeps the states.

[173] Khatua, A., Cambria, E., and Khatua, A. (2018). Sounds of silence breakers: Exploring sexual violence on twitter. In *2018 IEEE/ACM Int'l Conf. on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 397–400. IEEE.

[174] Kiekens, W. J., Baams, L., Fish, J. N., and Watson, R. J. (2021). Associations of Relationship Experiences, Dating Violence, Sexual Harassment, and Assault With Alcohol Use Among Sexual and Gender Minority Adolescents. *Journal of Interpersonal Violence*, page 08862605211001469. Publisher: SAGE Publications Inc.

[175] Kieslich, K., Keller, B., and Starke, C. (2022). Artificial intelligence ethics by design. evaluating public perception on the importance of ethical design principles of artificial intelligence. *Big Data & Society*, 9(1):20539517221092956.

[176] Kim, A. and Yoon, S. (2022). Detecting rumor veracity with only textual information by double-channel structure.

[177] Kim, B. K., Park, J., Jung, H. J., and Han, Y. (2020). Latent profiles of offline/cyber bullying experiences among korean students and its relationship with peer conformity. *Children and Youth Services Review*, 118:105349.

[178] Kim, S., Razi, A., Alsoubai, A., Ling, C., Stringhini, G., Wisniewski, P. J., and De Choudhury, M. (2022). I'm talking to you: Detecting and differentiating between online harassment in networked public vs. private social media spaces.

[179] Kim, S., Razi, A., Stringhini, G., Wisniewski, P. J., and De Choudhury, M. (2021a). A human-centered systematic literature review of cyberbullying detection algorithms. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–34.

[180] Kim, S., Razi, A., Stringhini, G., Wisniewski, P. J., and De Choudhury, M. (2021b). 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*, volume 15, pages 290–302.

[181] Kim, Y. (2014). Convolutional neural networks for sentence classification.

[182] Kirchner, T., Magallón-Neri, E., Forns, M., Muñoz, D., Segura, A., Soler, L., and Planellas, I. (2020). Facing interpersonal violence: identifying the coping profile of poly-victimized resilient adolescents. *Journal of interpersonal violence*, 35(9-10):1934–1957.

[183] Klonsky, E. D. and Glenn, C. R. (2009). Assessing the functions of non-suicidal self-injury: Psychometric properties of the inventory of statements about self-injury (isas). *Journal of psychopathology and behavioral assessment*, 31(3):215–219.

[184] Knijnenburg, B. P., Kobsa, A., and Jin, H. (2013). Dimensionality of information disclosure behavior. *International Journal of Human-Computer Studies*, 71(12):1144–1162.

[185] Koenen, K., Ratanatharathorn, A., Ng, L., McLaughlin, K., Bromet, E., Stein, D., Karam, E. G., Ruscio, A. M., Benjet, C., Scott, K., et al. (2017). Posttraumatic stress disorder in the world mental health surveys. *Psychological medicine*, 47(13):2260–2274.

[186] Korenis, P. and Billick, S. B. (2014). Forensic implications: Adolescent sexting and cyberbullying. *Psychiatric quarterly*, 85(1):97–101.

[187] Kozyreva, A., Herzog, S. M., Lewandowsky, S., Hertwig, R., Lorenz-Spreen, P., Leiser, M., and Reifler, J. (2023). Resolving content moderation dilemmas between free speech and harmful misinformation. *Proceedings of the National Academy of Sciences*, 120(7):e2210666120.

[188] Laborde, S. (2023). Teenage social media usage statistics in 2023.

[189] Laffey, T. J., Cox, P. A., Schmidt, J. L., Kao, S. M., and Readk, J. Y. (1988). Real-time knowledge-based systems. *AI magazine*, 9(1):27–27.

[190] Leiva, V. and Freire, A. (2017). Towards suicide prevention: early detection of depression on social media. In *Internet Science: 4th International Conference, INSCI 2017, Thessaloniki, Greece, November 22-24, 2017, Proceedings 4*, pages 428–436. Springer.

[191] Lenhart, A. (2009). *Teens and sexting*, volume 15. Pew Internet & American Life Project Washington, DC.

[192] Li, G., Gomez, R., Nakamura, K., and He, B. (2019). Human-centered reinforcement learning: A survey. *IEEE Transactions on Human-Machine Systems*, 49(4):337–349.

[193] Li, K., Guo, B., Liu, J., Wang, J., Ren, H., Yi, F., and Yu, Z. (2022a). Dynamic probabilistic graphical model for progressive fake news detection on social media platform. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 13(5):1–24.

[194] Li, K., Guo, B., Ren, S., and Yu, Z. (2022b). Adadebunk: An efficient and reliable deep state space model for adaptive fake news early detection. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 1156–1165.

[195] Li, X.-L. and Liu, B. (2005). Learning from positive and unlabeled examples with different data distributions. In *Machine Learning: ECML 2005: 16th European Conference on Machine Learning, Porto, Portugal, October 3-7, 2005. Proceedings 16*, pages 218–229. Springer.

[196] Li, Y. (2017). Deep reinforcement learning: An overview. *arXiv preprint arXiv:1701.07274*.

[197] Lin, K.-Y., Lee, R. K.-W., Gao, W., and Peng, W.-C. (2021). Early prediction of hate speech propagation. In *2021 International Conference on Data Mining Workshops (ICDMW)*, pages 967–974. IEEE.

[198] Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

[199] Links, P. S., Eynan, R., Heisel, M. J., and Nisenbaum, R. (2008). Elements of affective instability associated with suicidal behaviour in patients with borderline personality disorder. *The Canadian Journal of Psychiatry*, 53(2):112–116.

[200] Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57.

[201] Litou, I., Kalogeraki, V., Katakis, I., and Gunopulos, D. (2016). Real-time and cost-effective limitation of misinformation propagation. In *2016 17th IEEE International Conference on Mobile Data Management (MDM)*, volume 1, pages 158–163. IEEE.

[202] Liu, P., Guberman, J., Hemphill, L., and Culotta, A. (2018). Forecasting the presence and intensity of hostility on instagram using linguistic and social features. In *Twelfth international aaai conference on web and social media*.

[203] Liu, X., Nourbakhsh, A., Li, Q., Fang, R., and Shah, S. (2015). Real-time rumor debunking on twitter. In *Proceedings of the 24th ACM international on conference on information and knowledge management*, pages 1867–1870.

[204] Liu, Y., Li, Q., Liu, X., Zhang, Q., and Si, L. (2019). Sexual harassment story classification and key information identification. In *Proceedings of the 28th ACM Int'l Conf. on Information and Knowledge Management*, pages 2385–2388.

[205] Liu, Y. and Wu, Y.-F. (2018). Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

[206] Liu, Y. and Wu, Y.-F. B. (2020). Fned: a deep network for fake news early detection on social media. *ACM Transactions on Information Systems (TOIS)*, 38(3):1–33.

[207] Livingstone, S. and Helsper, E. (2010). Balancing opportunities and risks in teenagers' use of the internet: the role of online skills and internet self-efficacy. *New Media & Society*, 12(2):309–329.

[208] López-Vizcaíno, M. F., Nóvoa, F. J., Carneiro, V., and Cacheda, F. (2021). Early detection of cyberbullying on social media networks. *Future Generation Computer Systems*, 118:219–229.

[209] Losada, D. E., Crestani, F., and Parapar, J. (2017). erisk 2017: Clef lab on early risk prediction on the internet: experimental foundations. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 8th International Conference of the CLEF Association, CLEF 2017, Dublin, Ireland, September 11–14, 2017, Proceedings 8*, pages 346–360. Springer.

[210] Losada, D. E., Crestani, F., and Parapar, J. (2018). Overview of erisk: early risk prediction on the internet. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 9th International Conference of the CLEF Association, CLEF 2018, Avignon, France, September 10-14, 2018, Proceedings 9*, pages 343–361. Springer.

[211] Losada, D. E., Crestani, F., and Parapar, J. (2019). Overview of erisk at clef 2019: Early risk prediction on the internet (extended overview). *CLEF (Working Notes)*.

[212] Losada, D. E., Crestani, F., and Parapar, J. (2020). Overview of erisk at clef 2020: Early risk prediction on the internet (extended overview). *CLEF (Working Notes)*.

[213] Lu, M., Huang, Z., Li, B., Zhao, Y., Qin, Z., and Li, D. (2022). Sifter: A framework for robust rumor detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:429–442.

[214] Lubke, G. H. and Muthén, B. (2005). Investigating population heterogeneity with factor mixture models. *Psychological methods*, 10(1):21.

[215] Luo, Z., Sun, T., Zhu, X., Qian, Z., and Li, P. (2021). Early rumor detection with prior information on social media. In *Neural Information Processing: 28th International Conference, ICONIP 2021, Sanur, Bali, Indonesia, December 8–12, 2021, Proceedings, Part V 28*, pages 282–289. Springer.

[216] Lysyanskaya, A., Rivest, R. L., Sahai, A., and Wolf, S. (1999). Pseudonym systems. In *Int'l Workshop on Selected Areas in Cryptography*, pages 184–199. Springer.

[217] Ma, J., Gao, W., Wei, Z., Lu, Y., and Wong, K.-F. (2015). Detect rumors using time series of social context information on microblogging websites. In *Proceedings of the 24th ACM international on conference on information and knowledge management*, pages 1751–1754.

[218] Macoun, S. J., Bedir, B., Runions, K., Barker, L. E., Halliday, D., and Lewis, J. (2021). Information & communication technologies use by children & youth with autism spectrum disorder: Promise and perils. *Journal of Psychiatry & Behavioral Sciences*, 4(1):1047–1056.

[219] Madigan, S., Villani, V., Azzopardi, C., Laut, D., Smith, T., Temple, J. R., Browne, D., and Dimitropoulos, G. (2018). The prevalence of unwanted online sexual exposure and solicitation among youth: a meta-analysis. *Journal of Adolescent Health*, 63(2):133–141.

[220] Malhotra, A. and Jindal, R. (2022). Deep learning techniques for suicide and depression detection from online social media: A scoping review. *Applied Soft Computing*, page 109713.

[221] Mannekote Thippaiah, S., Shankarapura Nanjappa, M., Gude, J. G., Voyiaziakis, E., Patwa, S., Birur, B., and Pandurangi, A. (2021). Non-suicidal self-injury in developing countries: A review. *International journal of social psychiatry*, 67(5):472–482.

[222] Masaki, H., Shibata, K., Hoshino, S., Ishihama, T., Saito, N., and Yatani, K. (2020). Exploring nudge designs to help adolescent sns users avoid privacy and safety threats. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–11.

[223] Masud, S., Bedi, M., Khan, M. A., Akhtar, M. S., and Chakraborty, T. (2022). Proactively reducing the hate intensity of online posts via hate speech normalization. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3524–3534.

[224] McElvaney, R., Moore, K., O'Reilly, K., Turner, R., Walsh, B., and Guerin, S. (2020). Child sexual abuse disclosures: Does age make a difference? *Child abuse & neglect*, 99:104121.

[225] McHugh, C. M., Ho, N., Iorfino, F., Crouse, J. J., Nichles, A., Zmicerevska, N., Scott, E., Glozier, N., and Hickie, I. B. (2023). Predictive modelling of deliberate self-harm and suicide attempts in young people accessing primary care: a machine learning analysis of a longitudinal study. *Social psychiatry and psychiatric epidemiology*, pages 1–13.

[226] McHugh, M. L. (2011). Multiple comparison analysis testing in anova. *Biochemia medica*, 21(3):203–209.

[227] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35.

[228] Melton, G. B. (2005). Mandated reporting: a policy without reason. *Child abuse & neglect*, 29(1):9–18.

[229] Memon, A. M., Sharma, S. G., Mohite, S. S., and Jain, S. (2018). The role of online social networking on deliberate self-harm and suicidality in adolescents: A systematized review of literature. *Indian journal of psychiatry*, 60(4):384.

[230] Mia, V. (2020). The failures of sesta/fosta: A sex worker manifesto. *Transgender Studies Quarterly*, 7(2):237–239.

[231] Mimno, D., Wallach, H., Talley, E., Leenders, M., and McCallum, A. (2011). Optimizing semantic coherence in topic models. In *Proceedings of the 2011 Conf. on empirical methods in natural language processing*, pages 262–272.

[232] Min, E., Rong, Y., Bian, Y., Xu, T., Zhao, P., Huang, J., and Ananiadou, S. (2022). Divide-and-conquer: Post-user interaction network for fake news detection on social media. In *Proceedings of the ACM Web Conference 2022*, pages 1148–1158.

[233] Misra, K., Devarapalli, H., Ringenberg, T. R., and Rayz, J. T. (2019). Authorship analysis of online predatory conversations using character level convolution neural networks. In *2019 IEEE Int'l Conf. on Systems, Man and Cybernetics (SMC)*, pages 623–628. IEEE.

[234] Mitchell, K. J., Finkelhor, D., and Wolak, J. (2005). The internet and family and acquaintance sexual abuse. *Child maltreatment*, 10(1):49–60.

[235] Mitchell, K. J. and Jones, L. M. (2011). Youth internet safety study (yiss): Methodology report.

[236] Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., and Kavukcuoglu, K. (2016). Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pages 1928–1937. PMLR.

[237] Mnih, V., Heess, N., Graves, A., et al. (2014). Recurrent models of visual attention. *Advances in neural information processing systems*, 27.

[238] Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M. (2013). Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*.

[239] Monarch, R. M. (2021). *Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI*. Simon and Schuster.

[240] Morelli, M., Bianchi, D., Baiocco, R., Pezzuti, L., and Chirumbolo, A. (2016). Sexting, psychological distress and dating violence among adolescents and young adults. *Psicothema*, pages 137–142.

[241] Moreno, M. A., Jolliff, A., and Kerr, B. (2021). Youth advisory boards: Perspectives and processes. *Journal of Adolescent Health*, 69(2):192–194.

[242] Mori, C., Temple, J. R., Browne, D., and Madigan, S. (2019). Association of sexting with sexual behaviors and mental health among adolescents: A systematic review and meta-analysis. *JAMA pediatrics*, 173(8):770–779.

[243] Murphy, D. M. and Spencer, B. (2021). Teens' experiences with sexting: A grounded theory study. *Journal of Pediatric Health Care*.

[244] Muthén, B. and Muthén, B. O. (2009). *Statistical analysis with latent variables*. Wiley New York, NY.

[245] Naezer, M. (2018). From risky behaviour to sexy adventures: Reconceptualising young people's online sexual activities. *Culture, Health & Sexuality*, 20(6):715–729.

[246] Naumzik, C. and Feuerriegel, S. (2022). Detecting false rumors from retweet dynamics on social media. In *Proceedings of the ACM Web Conference 2022*, pages 2798–2809.

[247] Nazar, I., Zois, D.-S., and Yao, M. (2019). A hierarchical approach for timely cyberbullying detection. In *2019 IEEE Data Science Workshop (DSW)*, pages 190–195. IEEE.

[248] Nesi, J. (2020). The impact of social media on youth mental health: challenges and opportunities. *North Carolina medical journal*, 81(2):116–121.

[249] Newman, E., Risch, E., and Kassam-Adams, N. (2006). Ethical issues in trauma-related research: A review. *Journal of Empirical Research on Human Research Ethics*, 1(3):29–46.

[250] Nilizadeh, S., Labrèche, F., Sedighian, A., Zand, A., Fernandez, J., Kruegel, C., Stringhini, G., and Vigna, G. (2017). Poised: Spotting twitter spam off the beaten paths. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 1159–1174.

[251] Nishihara, R., Moritz, P., Wang, S., Tumanov, A., Paul, W., Schleier-Smith, J., Liaw, R., Niknami, M., Jordan, M. I., and Stoica, I. (2017). Real-time machine learning: The missing pieces. In *Proceedings of the 16th Workshop on Hot Topics in Operating Systems*, pages 106–110.

[252] Nosowsky, R. and Giordano, T. J. (2006). The health insurance portability and accountability act of 1996 (hipaa) privacy rule: implications for clinical research. *Annu. Rev. Med.*, 57:575–590.

[253] Nour, M. M., Rouf, A. S., and Allman-Farinelli, M. (2018). Exploring young adult perspectives on the use of gamification and social media in a smartphone platform for improving vegetable intake. *Appetite*, 120:547–556.

[254] Nova, F. F., Rifat, M. R., Saha, P., Ahmed, S. I., and Guha, S. (2019). Online sexual harassment over anonymous social media in bangladesh. In *Proceedings of the Tenth International Conference on Information and Communication Technologies and Development*, pages 1–12.

[255] Nylund, K. L., Asparouhov, T., and Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A monte carlo simulation study. *Structural equation modeling: A multidisciplinary Journal*, 14(4):535–569.

[256] Obaidat, I., Al-zou'bi, A., Mughaid, A., and Abualigah, L. (2023). Investigating the cyberbullying risk in digital media: protecting victims in school teenagers. *Social Network Analysis and Mining*, 13(1):139.

[of Economics and Science] of Economics, L. S. and Science, P. Euconsent - a child-rights approach to online age assurance and parental consent solutions.

[258] of Justice Programs, O. (2023). Internet safety: Online safety for youth.

[259] Oliva, T. D., Antonialli, D. M., and Gomes, A. (2021). Fighting hate speech, silencing drag queens? artificial intelligence in content moderation and risks to lgbtq voices online. *Sexuality & Culture*, 25(2):700–732.

[260] O'Reilly, C. A. and Cromarty, A. S. (1985). " fast" is not" real-time": Designing effective real-time ai systems. In *Applications of Artificial Intelligence II*, volume 548, pages 249–257. SPIE.

[261] O'reilly, M. (2020). Social media and adolescent mental health: the good, the bad and the ugly. *Journal of Mental Health*, 29(2):200–206.

[262] Pabian, S., Erreygers, S., Vandebosch, H., Van Royen, K., Dare, J., Costello, L., Green, L., Hawk, D., and Cross, D. (2018). "arguments online, but in school we always act normal": The embeddedness of early adolescent negative peer interactions within the whole of their offline and online peer interactions. *Children and youth services review*, 86:1–13.

[263] Palen, L. and Dourish, P. (2003). Unpacking" privacy" for a networked world. In *Proceedings of the SIGCHI Conf. on Human factors in computing systems*, pages 129–136.

[264] Palmier-Claus, J., Taylor, P. J., Varese, F., and Pratt, D. (2012). Does unstable mood increase risk of suicide? theory, research and practice. *Journal of affective disorders*, 143(1-3):5–15.

[265] Parapar, J., Martín-Rodilla, P., Losada, D. E., and Crestani, F. (2021). Overview of erisk at clef 2021: Early risk prediction on the internet (extended overview). *CLEF (Working Notes)*, pages 864–887.

[266] Parapar, J., Martín-Rodilla, P., Losada, D. E., and Crestani, F. (2022). Overview of erisk 2022: Early risk prediction on the internet. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 13th International Conference of the CLEF Association, CLEF 2022, Bologna, Italy, September 5–8, 2022, Proceedings*, pages 233–256. Springer.

[267] Parapar, J., Martín-Rodilla, P., Losada, D. E., and Crestani, F. (2023). Overview of erisk 2023: Early risk prediction on the internet. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 294–315. Springer.

[268] Park, J., Arunachalam, R., Silenzio, V., Singh, V. K., et al. (2022). Fairness in mobile phone–based mental health assessment algorithms: Exploratory study. *JMIR formative research*, 6(6):e34366.

[269] Park, J., Ellezhuthil, R. D., Isaac, J., Mergerson, C., Feldman, L., and Singh, V. (2023a). Misinformation detection algorithms and fairness across political ideologies: The impact of article level labeling. In *Proceedings of the 15th ACM Web Science Conference 2023*, pages 107–116.

[270] Park, J., Gracie, J., Alsoubai, A., Stringhini, G., Singh, V., and Wisniewski, P. (2023b). Towards automated detection of risky images shared by youth on social media. In *Companion Proceedings of the ACM Web Conference 2023*, pages 1348–1357.

[271] Patchin, J. W. and Hinduja, S. (2020). It is time to teach safe sexting. *Journal of Adolescent Health*, 66(2):140–143.

[272] Patil, S., Gune, A., and Nene, M. (2017). Convolutional neural networks for text categorization with latent semantic analysis. In *2017 Int'l Conf. on Energy, Communication, Data Analytics and Soft Computing (ICECDS)*, pages 499–503. IEEE.

[273] Petrescu, A., Truică, C.-O., Apostol, E.-S., and Karras, P. (2021). Sparse shield: Social network immunization vs. harmful speech. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 1426–1436.

[274] Phillips, L., Dowling, C., Shaffer, K., Hodas, N., and Volkova, S. (2017). Using social media to predict the future: a systematic literature review. *arXiv preprint arXiv:1706.06134*.

[275] Pichel, R., Feijóo, S., Isorna, M., Varela, J., and Rial, A. (2022). Analysis of the relationship between school bullying, cyberbullying, and substance use. *Children and Youth Services Review*, 134:106369.

[276] Pinter, A. T., Wisniewski, P. J., Xu, H., Rosson, M. B., and Caroll, J. M. (2017). Adolescent online safety: Moving beyond formative evaluations to designing solutions for the future. In *Proceedings of the 2017 Conference on Interaction Design and Children*, pages 352–357.

[277] Ponton, L. E. and Judice, S. (2004). Typical adolescent sexual development. *Child and Adolescent Psychiatric Clinics*, 13(3):497–511.

[278] Post, N. Y. (2023). China is hurting our kids with tiktok but protecting its own youth with douyin.

[279] Potha, N. and Maragoudakis, M. (2014). Cyberbullying detection using time series modeling. In *2014 IEEE International Conference on Data Mining Workshop*, pages 373–382. IEEE.

[280] Prino, L. E., Longobardi, C., and Settanni, M. (2018). Young adult retrospective reports of adverse childhood experiences: Prevalence of physical, emotional, and sexual abuse in italy. *Archives of sexual behavior*, 47(6):1769–1778.

[281] Radesky, J. and Hiniker, A. (2021). From moral panic to systemic change: Making child-centered design the default. *Int'l Journal of Child-Computer Interaction*, page 100351.

[282] Rafiq, R. I., Hosseinmardi, H., Han, R., Lv, Q., and Mishra, S. (2018). Scalable and timely detection of cyberbullying in online social networks. In *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*, pages 1738–1747.

[283] Ramírez-Cifuentes, D., Mayans, M., and Freire, A. (2018). Early risk detection of anorexia on social media. In *Internet Science: 5th International Conference, INSCI 2018, St. Petersburg, Russia, October 24–26, 2018, Proceedings 5*, pages 3–14. Springer.

[284] Rastogi, S. and Bansal, D. (2022). A review on fake news detection 3t's: typology, time of detection, taxonomies. *International Journal of Information Security*, pages 1–36.

[285] Razi, A., AlSoubai, A., Kim, S., Ali, S., Stringhini, G., De Choudhury, M., and Wisniewski, P. J. (2022a). Sliding into my dms: Detecting uncomfortable or unsafe sexual risk experiences within instagram direct messages grounded in the perspective of youth.

[286] Razi, A., AlSoubai, A., Kim, S., Ali, S., Stringhini, G., De Choudhury, M., and Wisniewski, P. J. (2023). Sliding into my dms: Detecting uncomfortable or unsafe sexual risk experiences within instagram direct messages grounded in the perspective of youth. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):1–29.

[287] Razi, A., AlSoubai, A., Kim, S., Naher, N., Ali, S., Stringhini, G., De Choudhury, M., and Wisniewski, P. J. (2022b). Instagram data donation: A case study on collecting ecologically valid social media data for the purpose of adolescent online risk detection. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, pages 1–9.

[288] Razi, A., Badillo-Urquiola, K., and Wisniewski, P. J. (2020). Let's talk about sext: How adolescents seek support and advice about their online sexual experiences. In *Proceedings of the 2020 CHI Conf. on Human Factors in Computing Systems*, pages 1–13.

[289] Razi, A., Kim, S., Alsoubai, A., Caddle, X., Ali, S., Stringhini, G., Choudhury, M. D., and Wisniewski, P. (2021a). 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*.

[290] Razi, A., Kim, S., Alsoubai, A., Stringhini, G., Solorio, T., De Choudhury, M., and Wisniewski, P. J. (2021b). A human-centered systematic literature review of the computational approaches for online sexual risk detection. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–38.

[291] Regulation, G. D. P. (2016). Regulation eu 2016/679 of the european parliament and of the council of 27 april 2016. *Official Journal of the European Union*.

[292] Reid, D. and Weigle, P. (2014). Social media use among adolescents: Benefits and risks. *Adolescent Psychiatry*, 4(2):73–80.

[293] Ríos Vega, J. A. (2020). School to deportation pipeline: Latino youth counter-storytelling narratives. *Journal of Latinos and Education*, pages 1–13.

[294] Rodrigues, L., Toda, A. M., Palomino, P. T., Oliveira, W., and Isotani, S. (2020). Personalized gamification: A literature review of outcomes, experiments, and approaches. In *Eighth international conference on technological ecosystems for enhancing multiculturality*, pages 699–706.

[295] Rodríguez-Enríquez, M., Bennasar-Veny, M., Leiva, A., and Yañez, A. M. (2019). Alcohol and tobacco consumption, personality, and cybervictimization among adolescents. *International journal of environmental research and public health*, 16(17):3123.

[296] Rogers, D., Preece, A., Innes, M., and Spasić, I. (2022). Real-time text classification of user-generated content on social media: Systematic review. *IEEE Transactions on Computational Social Systems*, 9(4):1154–1166.

[297] Rosenfeld, N., Szanto, A., and Parkes, D. C. (2020). A kernel of truth: Determining rumor veracity on twitter by diffusion pattern alone. In *Proceedings of The Web Conference 2020*, pages 1018–1028.

[298] Rothchild, J. A. (2016). *Research handbook on electronic commerce law*. Edward Elgar Publishing.

[299] Rousidis, D., Koukaras, P., and Tjortjis, C. (2020). Social media prediction: a literature review. *Multimedia Tools and Applications*, 79(9-10):6279–6311.

[300] Salter, M. and Crofts, T. (2015). Responding to revenge porn: Challenges to online legal impunity. *New views on pornography: Sexuality, politics, and the law*, pages 233–256.

[301] Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., and Müller, K.-R. (2019). *Explainable AI: interpreting, explaining and visualizing deep learning*, volume 11700. Springer Nature.

[302] Samuels-Wortley, K. (2021). To serve and protect whom? using composite counter-storytelling to explore black and indigenous youth experiences and perceptions of the police in canada. *Crime & Delinquency*, 67(8):1137–1164.

[303] Sandvig, C., Hamilton, K., Karahalios, K., and Langbort, C. (2014). Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and discrimination: converting critical concerns into productive inquiry*, 22(2014):4349–4357.

[304] Santhoshkumar, S. and Dhinesh Babu, L. (2020). Earlier detection of rumors in online social networks using certainty-factor-based convolutional neural networks. *Social Network Analysis and Mining*, 10:1–17.

[305] Sawhney, R., Agarwal, S., Neerkaje, A. T., Aletras, N., Nakov, P., and Flek, L. (2022). Towards suicide ideation detection through online conversational context. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, pages 1716–1727.

[306] Sawhney, R., Joshi, H., Gandhi, S., and Shah, R. (2020). A time-aware transformer based model for suicide ideation detection on social media. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 7685–7697.

[307] Sawhney, R., Joshi, H., Shah, R., and Flek, L. (2021). Suicide ideation detection via social and temporal user representations using hyperbolic learning. In *Proceedings of the 2021 conference of the North American Chapter of the Association for Computational Linguistics: human language technologies*, pages 2176–2190.

[308] Saxena, D., Badillo-Urquiola, K., Wisniewski, P. J., and Guha, S. (2020). 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.

[309] Seidel, V. P., Hannigan, T. R., and Phillips, N. (2020). Rumor communities, social media, and forthcoming innovations: The shaping of technological frames in product market evolution. *Academy of Management Review*, 45(2):304–324.

[310] Seliya, N., Khoshgoftaar, T. M., and Van Hulse, J. (2009). A study on the relationships of classifier performance metrics. In *2009 21st IEEE international conference on tools with artificial intelligence*, pages 59–66. IEEE.

[311] Serafini, G., Aguglia, A., Amerio, A., Canepa, G., Adavastro, G., Conigliaro, C., Nebbia, J., Franchi, L., Flouri, E., and Amore, M. (2021). The relationship between bullying victimization and perpetration and non-suicidal self-injury: a systematic review. *Child Psychiatry & Human Development*, pages 1–22.

[312] Sevtiyun, P. E., Oktadini, N. R., and Bardadi, A. (2020). Information risk assessment model of accuracy and timeliness dimensions. In *Sriwijaya International Conference on Information Technology and Its Applications (SICONIAN 2019)*, pages 634–639. Atlantis Press.

[313] Shalhoub-Kevorkian, N. (2005). Disclosure of child abuse in conflict areas. *Violence Against Women*, 11(10):1263–1291.

[314] Shannon, C. E. (2001). A mathematical theory of communication. *ACM SIGMOBILE mobile computing and communications review*, 5(1):3–55.

[315] Shapka, J. D. and Maghsoudi, R. (2017). Examining the validity and reliability of the cyber-aggression and cyber-victimization scale. *Computers in Human Behavior*, 69:10–17.

[316] Sharpe, D. (2015). Chi-square test is statistically significant: Now what? *Practical Assessment, Research, and Evaluation*, 20(1):8.

[317] Shin, K. G. and Ramanathan, P. (1994). Real-time computing: A new discipline of computer science and engineering. *Proceedings of the IEEE*, 82(1):6–24.

[318] Shiryaev, A. N. (2007). *Optimal stopping rules*, volume 8. Springer Science & Business Media.

[319] Shneiderman, B. (2020). Bridging the gap between ethics and practice: guidelines for reliable, safe, and trustworthy human-centered ai systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 10(4):1–31.

[320] Shneiderman, B., Plaisant, C., Cohen, M. S., Jacobs, S., Elmqvist, N., and Diakopoulos, N. (2016). *Designing the user interface: strategies for effective human-computer interaction*. Pearson.

[321] Shu, K., Sliva, A., Wang, S., Tang, J., and Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36.

[322] Silva, A., Han, Y., Luo, L., Karunasekera, S., and Leckie, C. (2021). Propagation2vec: Embedding partial propagation networks for explainable fake news early detection. *Information Processing & Management*, 58(5):102618.

[323] Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. (2016). Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489.

[324] Sinaga, K. P. and Yang, M.-S. (2020). Unsupervised k-means clustering algorithm. *IEEE access*, 8:80716–80727.

[325] Smith, C. L., Rhoades Cooper, B., Miguel, A., Roll, J., Hill, L., Cleveland, M., and McPherson, S. (2022). Youth risk profiles and their prediction of distal cannabis and tobacco co-use in the population assessment of tobacco health (path). *Substance Abuse*, 43(1):733–741.

[326] Sood, S. O., Antin, J., and Churchill, E. (2012). Using crowdsourcing to improve profanity detection. In *2012 AAAI Spring Symposium Series*.

[327] Starke, C., Baleis, J., Keller, B., and Marcinkowski, F. (2022). Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature. *Big Data & Society*, 9(2):20539517221115189.

[328] Subrahmanian, V. and Kumar, S. (2017). Predicting human behavior: The next frontiers. *Science*, 355(6324):489–489.

[329] Subrahmanyam, K. and Smahel, D. (2010). *Digital youth: The role of media in development*. Springer Science & Business Media.

[330] Suresh, H., Gomez, S. R., Nam, K. K., and Satyanarayan, A. (2021). Beyond expertise and roles: A framework to characterize the stakeholders of interpretable machine learning and their needs. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–16.

[331] Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

[332] Suzuki, L. K. and Calzo, J. P. (2004). 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.

[333] Tariq, M. U., Razi, A., Badillo-Urquiola, K., and Wisniewski, P. (2019). A review of the gaps and opportunities of nudity and skin detection algorithmic research for the purpose of combating adolescent sexting behaviors. In *Human-Computer Interaction. Design Practice in Contemporary Societies: Thematic Area, HCI 2019, Held as Part of the 21st HCI International Conference, HCII 2019, Orlando, FL, USA, July 26–31, 2019, Proceedings, Part III 21*, pages 90–108. Springer.

[334] Tarrant, M. A., Manfredo, M. J., Bayley, P. B., and Hess, R. (1993). Effects of recall bias and nonresponse bias on self-report estimates of angling participation. *North American Journal of Fisheries Management*, 13(2):217–222.

[335] Tarrier, N., Gooding, P., Gregg, L., Johnson, J., Drake, R., Group, S. T., et al. (2007). Suicide schema in schizophrenia: The effect of emotional reactivity, negative symptoms and schema elaboration. *Behaviour research and therapy*, 45(9):2090–2097.

[Team] Team, T. B. Teen text speak codes every parent should know.

[337] Technologies, B. (2023). Bark technologies releases 2022 annual report.

[338] Temple, J. R., Le, V. D., van den Berg, P., Ling, Y., Paul, J. A., and Temple, B. W. (2014). Brief report: Teen sexting and psychosocial health. *Journal of adolescence*, 37(1):33–36.

[339] Thieme, A., Belgrave, D., and Doherty, G. (2020). Machine learning in mental health: A systematic review of the hci literature to support the development of effective and implementable ml systems. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 27(5):1–53.

[340] Tolman, D. L. and McClelland, S. I. (2011). Normative sexuality development in adolescence: A decade in review, 2000–2009. *Journal of research on adolescence*, 21(1):242–255.

[341] Ullman, S. E. (2007). Relationship to perpetrator, disclosure, social reactions, and ptsd symptoms in child sexual abuse survivors. *Journal of child sexual abuse*, 16(1):19–36.

[342] Vaccaro, K., Sandvig, C., and Karahalios, K. (2020). " at the end of the day facebook does what itwants" how users experience contesting algorithmic content moderation. *Proceedings of the ACM on human-computer interaction*, 4(CSCW2):1–22.

[343] Valiukas, S., Pickering, M., Hall, T., Seneviratne, N., Aitken, A., John-Leader, F., and Pit, S. W. (2019). Sexting and mental health among young australians attending a musical festival: a cross sext-ional study. *Cyberpsychology, Behavior, and Social Networking*, 22(8):521–528.

[344] Van Ouytsel, J., Van Gool, E., Ponnet, K., and Walrave, M. (2014). Brief report: The association between adolescents' characteristics and engagement in sexting. *Journal of adolescence*, 37(8):1387–1391.

[345] Van Ouytsel, J., Walrave, M., De Marez, L., Vanhaelewyn, B., and Ponnet, K. (2020). A first investigation into gender minority adolescents' sexting experiences. *Journal of Adolescence*, 84:213–218.

[346] Van Ouytsel, J., Walrave, M., Ponnet, K., and Heirman, W. (2015). The association between adolescent sexting, psychosocial difficulties, and risk behavior: Integrative review. *The Journal of School Nursing*, 31(1):54–69.

[347] Vanden Abeele, M., Campbell, S. W., Eggermont, S., and Roe, K. (2014). Sexting, mobile porn use, and peer group dynamics: Boys' and girls' self-perceived popularity, need for popularity, and perceived peer pressure. *Media Psychology*, 17(1):6–33.

[348] Villacampa, C. (2017). Teen sexting: Prevalence, characteristics and legal treatment. *Int'l Journal of Law, Crime and Justice*, 49:10–21.

[349] Vogels, E. A. (2023). Teens and social media: Key findings from pew research center surveys.

[350] Vosoughi, S., Mohsenvand, M. and Roy, D. (2017). Rumor gauge: Predicting the veracity of rumors on twitter. *ACM transactions on knowledge discovery from data (TKDD)*, 11(4):1–36.

[351] Vyas, P., Vyas, G., Chauhan, A., Rawat, R., Telang, S., and Gottumukkala, M. (2022). Anonymous trading on the dark online marketplace: An exploratory study. In *Using Computational Intelligence for the Dark Web and Illicit Behavior Detection*, pages 272–289. IGI Global.

[Wachs] Wachs, S. Hate speech and bullying: Two sides of the same coin?

[353] Wachs, S., Wright, M. F., Gámez-Guadix, M., and Döring, N. (2021). How are consensual, non-consensual, and pressured sexting linked to depression and self-harm? the moderating effects of demographic variables. *International journal of environmental research and public health*, 18(5):2597.

[354] Walsh, J. P. (2020). Social media and moral panics: Assessing the effects of technological change on societal reaction. *International Journal of Cultural Studies*, 23(6):840–859.

[355] Walters, G. D. and Espelage, D. L. (2018). From victim to victimizer: Hostility, anger, and depression as mediators of the bullying victimization–bullying perpetration association. *Journal of school psychology*, 68:73–83.

[356] Wang, D., Yang, Q., Abdul, A., and Lim, B. Y. (2019). Designing theory-driven user-centric explainable ai. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–15.

[357] Wang, J., Zhao, P., Hoi, S. C., and Jin, R. (2013). Online feature selection and its applications. *IEEE Transactions on knowledge and data engineering*, 26(3):698–710.

[358] Wei, H., Kang, X., Wang, W., and Ying, L. (2019). Quickstop: A markov optimal stopping approach for quickest misinformation detection. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 3(2):1–25.

[359] White, S. F., Voss, J. L., Chiang, J. J., Wang, L., McLaughlin, K. A., and Miller, G. E. (2019). Exposure to violence and low family income are associated with heightened amygdala responsiveness to threat among adolescents. *Developmental cognitive neuroscience*, 40:100709.

[360] Wisniewski, P., Jia, H., Wang, N., Zheng, S., Xu, H., Rosson, M. B., and Carroll, J. M. (2015). Resilience mitigates the negative effects of adolescent internet addiction and online risk exposure. In *Proceedings of the 33rd Annual ACM Conf. on Human Factors in Computing Systems*, pages 4029–4038.

[361] Wisniewski, P., Xu, H., Rosson, M. B., and Carroll, J. M. (2017). Parents just don't understand: Why teens don't talk to parents about their online risk experiences. In *Proceedings of the 2017 ACM Conf. on computer supported cooperative work and social computing*, pages 523–540.

[362] Wisniewski, P., Xu, H., Rosson, M. B., Perkins, D. F., and Carroll, J. M. (2016a). Dear diary: Teens reflect on their weekly online risk experiences. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 3919–3930.

[363] Wisniewski, P., Xu, H., Rosson, M. B., Perkins, D. F., and Carroll, J. M. (2016b). 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. ACM. event-place: San Jose, California, USA.

[364] Wolf, C. T. (2019). Explainability scenarios: towards scenario-based xai design. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pages 252–257.

[365] Wrońska, A., Lew-Starowicz, R., and Rywczyńska, A. (2020). *Education–Relationships–Play Multi-faceted Aspects of the Internet and Child and Youth Online Safety*. Foundation for the Development of the Education System.

[366] Wu, C.-Y., Stewart, R., Huang, H.-C., Prince, M., and Liu, S.-I. (2011). The impact of quality and quantity of social support on help-seeking behavior prior to deliberate self-harm. *General hospital psychiatry*, 33(1):37–44.

[367] Wu, L., Li, J., Hu, X., and Liu, H. (2017). Gleaning wisdom from the past: Early detection of emerging rumors in social media. In *Proceedings of the 2017 SIAM international conference on data mining*, pages 99–107. SIAM.

[368] Wu, L. and Liu, H. (2019). Debunking rumors in social networks: A timely approach. In *Proceedings of the 10th ACM Conference on Web Science*, pages 323–331.

[369] Xia, R., Xuan, K., and Yu, J. (2020). A state-independent and time-evolving network for early rumor detection in social media. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 9042–9051.

[370] Xu, T., Goossen, G., Cevahir, H. K., Khodeir, S., Jin, Y., Li, F., Shan, S., Patel, S., Freeman, D., and Pearce, P. (2021). Deep entity classification: Abusive account detection for online social networks. In *30th {USENIX} Security Symposium ({USENIX} Security 21)*.

[371] Xu, X., Deng, K., and Zhang, X. (2022). Identifying cost-effective debunkers for multi-stage fake news mitigation campaigns. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pages 1206–1214.

[372] Xue, J., Zhang, B., Zhang, Q., Hu, R., Jiang, J., Liu, N., Peng, Y., Li, Z., and Logan, J. (2023). Using twitter-based data for sexual violence research: Scoping review. *Journal of Medical Internet Research*, 25:e46084.

[373] Yan, P., Li, L., Chen, W., and Zeng, D. (2019). Quantum-inspired density matrix encoder for sexual harassment personal stories classification. In *2019 IEEE Int'l Conf. on Intelligence and Security Informatics (ISI)*, pages 218–220. IEEE.

[374] Yao, M., Chelmis, C., and Zois, D.-S. (2019). Cyberbullying ends here: Towards robust detection of cyberbullying in social media. In *The World Wide Web Conference*, pages 3427–3433.

[375] Yazdavar, A. H., Al-Olimat, H. S., Ebrahimi, M., Bajaj, G., Banerjee, T., Thirunarayan, K., Pathak, J., and Sheth, A. (2017). Semi-supervised approach to monitoring clinical depressive symptoms in social media. In *Proceedings of the 2017 IEEE/ACM international conference on advances in social networks analysis and mining 2017*, pages 1191–1198.

[376] Ybarra, M. L. and Mitchell, K. J. (2014). “sexting” and its relation to sexual activity and sexual risk behavior in a national survey of adolescents. *Journal of adolescent health*, 55(6):757–764.

[377] Yi, P. and Zubiaga, A. (2022). Session-based cyberbullying detection in social media: A survey. *arXiv preprint arXiv:2207.10639*.

[378] Yi, P. and Zubiaga, A. (2023). Learning like human annotators: Cyberbullying detection in lengthy social media sessions. In *Proceedings of the ACM Web Conference 2023*, pages 4095–4103.

[379] Yin, J. and Wang, J. (2014). A dirichlet multinomial mixture model-based approach for short text clustering. *Proceedings of the 20th ACM SIGKDD Int’l Conf. on Knowledge discovery and data mining*.

[380] Yoon, Y., Lee, J. O., Cho, J., Bello, M. S., Khoddam, R., Riggs, N. R., and Leventhal, A. M. (2019). Association of cyberbullying involvement with subsequent substance use among adolescents. *Journal of Adolescent Health*, 65(5):613–620.

[381] Završnik, A. (2021). Algorithmic justice: Algorithms and big data in criminal justice settings. *European Journal of criminology*, 18(5):623–642.

[382] Zhang, J., Yamanaka, J., and Li, L. (2020). Early automatic detection of false information in twitter event considering occurrence scale and time series. In *Proceedings of the 22nd International Conference on Information Integration and Web-based Applications & Services*, pages 282–289.

[383] Zhao, T., Ni, B., Yu, W., Guo, Z., Shah, N., and Jiang, M. (2021). Action sequence augmentation for early graph-based anomaly detection. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 2668–2678.

[384] Zhou, K., Shu, C., Li, B., and Lau, J. H. (2019). Early rumour detection. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1614–1623.

[385] Zhou, X., Cao, J., Jin, Z., Xie, F., Su, Y., Chu, D., Cao, X., and Zhang, J. (2015). Real-time news certification system on sina weibo. In *Proceedings of the 24th International Conference on World Wide Web*, pages 983–988.

[386] Zhou, Y. (2020). A review of text classification based on deep learning. In *Proceedings of the 2020 3rd Int’l Conf. on Geoinformatics and Data Analysis*, pages 132–136.

[387] Zogan, H., Razzak, I., Jameel, S., and Xu, G. (2021). Depressionnet: A novel summarization boosted deep framework for depression detection on social media. *arXiv preprint arXiv:2105.10878*.

[Zuckerberg] Zuckerberg, M. New update for end-to-end encrypted messenger.

[389] Zych, I., Viejo, C., Vila, E., and Farrington, D. P. (2021). School bullying and dating violence in adolescents: A systematic review and meta-analysis. *Trauma, Violence, & Abuse*, 22(2):397–412.

[390] Zytko, D., Furlo, N., Carlin, B., and Archer, M. (2021). Computer-mediated consent to sex: The context of tinder. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):189:1–189:26.

Table 1: Social Media Platforms.

Social Media Platforms	Counts (Percent)	References
Twitter	36 (68%)	(58; 92; 98; 114; 197; 203; 232; 246; 250; 273; 297; 304; 306; 305; 322; 350; 367; 371; 375; 382; 387; 201; 307; 75; 205; 215; 217; 369; 384; 368; 223; 194; 193; 206; 213)
Weibo	14 (26%)	(75; 205; 215; 217; 369; 384; 194; 193; 206; 213; 159; 358; 385; 383)
Instagram	8 (15%)	(72; 77; 202; 247; 374; 78; 129; 378)
Vine	5 (9%)	(78; 129; 378; 208; 282)
Reddit	5 (9%)	(190; 283; 176; 223; 383)
Facebook	2 (4%)	(370; 223)

Table 2: Ground Truth Annotations

Annotations	Counts (Percent)	References
Existing	38 (72%)	(58; 72; 75; 78; 77; 98; 114; 129; 159; 176; 190; 194; 197; 201; 205; 206; 213; 215; 217; 232; 246; 247; 282; 283; 297; 304; 322; 358; 368; 369; 371; 374; 378; 382; 385; 384; 387; 223)
Outsiders	15 (28%)	(250; 203; 367; 370; 92; 306; 307; 193; 202; 208; 305; 350; 375; 383; 223)
Auto	5 (9%)	(273; 250; 203; 367; 370)

Table 3: Class Distribution

Class Distribution	Counts (Percent)	References
Unbalanced	31 (58%)	(58; 72; 78; 77; 98; 368; 114; 129; 176; 190; 197; 202; 208; 213; 215; 232; 247; 273; 283; 297; 304; 306; 307; 305; 322; 367; 370; 378; 383; 387; 206)
Balanced	22 (42%)	(75; 92; 159; 194; 193; 201; 203; 205; 217; 223; 246; 250; 282; 350; 358; 369; 371; 374; 375; 382; 385; 384)

Table 4: Dataset Processing

Dataset Type	Counts (Percent)	References
Chunks of data	44 (83%)	(75; 78; 92; 98; 114; 129; 159; 176; 194; 193; 197; 205; 202; 206; 208; 215; 217; 223; 232; 246; 250; 273; 282; 283; 297; 304; 306; 307; 305; 322; 350; 358; 367; 368; 369; 370; 371; 375; 378; 382; 383; 385; 384; 387; 58; 203)
Dynamical	9 (17%)	(58; 203; 72; 77; 190; 201; 213; 247; 374)

Table 5: Features

Features	**Counts (Percent)**	**Types**	**References**
ML-Based	53 (100%)	Textual (66%)	(159; 206; 72; 78; 77; 129; 283; 378; 382; 385; 213; 58; 75; 114; 176; 194; 205; 202; 215; 247; 307; 375; 383; 387; 190; 305; 369; 384; 306; 223; 193; 350; 197; 250; 374)
		Network (51%)	(201; 203; 358; 382; 387; 232; 273; 297; 371; 383; 217; 350; 385; 305; 197; 307; 322; 368; 205; 246; 304; 250; 98; 159; 129; 208; 78)
		User (30%)	(282; 217; 385; 205; 246; 304; 98; 159; 114; 206; 194; 370; 215; 129; 208; 78)
		Temporal (21%)	(197; 307; 322; 368; 205; 246; 304; 75; 369; 77; 78)
		Sentiment (19%)	(382; 387; 282; 217; 385; 72; 190; 223; 193; 208)
Domain-Specific	17 (32%)		(92; 176; 202; 367; 374; 375; 213; 282; 283; 201; 203; 358; 382; 387; 370; 194; 217)

Table 6: Machine Learning Models

Approach	**Counts (Percent)**	**Model**	**References**
Statistical	21 (40%)	Bayes	(194; 247; 193; 246; 374; 58; 250; 368; 208)
		Markov Models	(193; 350; 246; 369; 358; 72; 92; 190)
		Hawkes process	(307; 246; 371)
Deep Learning	31 (60%)	LSTM	(98; 223; 197; 273; 350; 305; 371; 306; 75; 194)
		Graph Neural Network	(297; 367; 322; 203; 383; 201; 307; 232)
		Transformers-Based	(176; 215; 387; 223; 273)
		CNN	(206; 304; 159)
		Neural Network	(114; 385; 213)
		Gated recurrent units	(383; 205)

Table 7: Models' Explainable Approaches

Models' Explanibility	**Counts (Percent)**	**References**
Qualitative analysis	16 (32%)	(202; 247; 322; 350; 306; 384; 305; 206; 58; 297; 387; 246; 213; 77; 378; 78)
Error analysis	7 (13%)	(223; 307; 250; 358; 58; 305; 367)
Case study	5 (9%)	(197; 193; 194; 383; 129)
Human evaluation	1 (2%)	(223)