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

Ashwaq Alsoubai
ashwaq.alsoubai@vanderbilt.edu
Vanderbilt University
Nashville, Tennessee, USA

ABSTRACT

Computational approaches to detect the online risks that the youth encounter have presented promising potentials to protect them online. However, a major identified trend among these approaches is the lack of human-centered machine learning (HCML) aspect. It is necessary to move beyond the computational lens of the detection task to address the societal needs of such a vulnerable population. Therefore, I direct my attention in this dissertation to better understand youths' risk experiences prior to enhancing the development of risk detection algorithms by 1) Examining youths' (ages 13–17) *public* disclosures about sexual experiences and contextualizing these experiences based on the levels of consent (i.e., consensual, non-consensual, sexual abuse) and relationship types (i.e., stranger, dating/friend, family), 2) Moving beyond the sexual experiences to examine a broader array of risks within the *private* conversations of youth (N = 173) between 13 and 21 and contextualizing the dynamics of youth online and offline risks and the self-reports of risk experiences to the digital trace data, and 3) Building real-time machine learning models for risk detection by creating a contextualized framework. This dissertation provides a human-centered approach for improving automated real-time risk predictions that are derived from a contextualized understanding of the nuances relative to youths' risk experiences.

ACM Reference Format:
Ashwaq Alsoubai. 2023. A Human-Centered Approach to Improving Adolescent Real-Time Online Risk Detection Algorithms. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (CHI EA '23), April 23–28, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3544549.3577045

1 INTRODUCTION

Most youths' social and developmental growth has been mediated through social media extensive usage [2]. 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 [30]. 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 [4, 18, 28]. 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 [39]. 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 [24].

Upon reviewing the current machine learning risk detection approaches incorporating the 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) [40], 2) lack youths' risk perspectives when identifying ground truth annotations [27], 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 [40]. 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 [1]. Therefore, in my dissertation, HCML will be applied primarily by deeply understanding the prevalent risks that youth encounter as well as the dynamics of these risks across contexts (e.g., online and offline), and then identifying an appropriate algorithmic contextual framework to advance 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:** *What insights can we gain regarding the online risk experiences of youth through their disclosures of sexual risk experiences when seeking peer-support online?*
- **RQ2:** *How does creating profiles of youth based on their self-reported online and offline risk behaviors inform their lived risk experiences on social media??*
- **RQ3:** *Can the insights gained from the prior two studies be built upon to create algorithms that accurately detect the most common risks youth encounter online in real-time?*

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)). The first study addressed RQ1 and was published at the ACM Conference On Computer-Supported Cooperative Work And Social Computing (CSCW 2022) [6]. The second study addressed RQ2 and was rejected from the Conference on Human Factors in Computing Systems (CHI 2023) for not citing enough papers to frame the paper clearly, which was something missing due to the

CHI EA '23, April 23–28, 2023, Hamburg, Germany

ACM ISBN 978-1-4503-9422-2/23/04.
https://doi.org/10.1145/3544549.3577045

word count limit; therefore, I planned to work with my advisor Dr. Pamela Wisnwiski to revise and resubmit it to CSCW (January 15). I also planned to work on study 3 to address RQ3 during 2023.

This year (Fall 2022 and Spring 2023) is my fourth year in the Ph.D. program of Computer Science at Vanderbilt and the plan is to graduate by the end of Spring 2024. The long-term goal of this dissertation is to incorporate the trained risk detection algorithms into a web-based dashboard for youths' evaluations. This dashboard leverages human-in-the-loop approach to enhance the accuracy of these models based on youth feedback on the predictions. Although this dashboard will not be part of my dissertation, I already built the web-system and published a demonstration about it in CSCW 2022 [5]. The three studies will be presented below in further detail.

2 FIRST STUDY (RQ1): FROM 'FRIENDS WITH BENEFITS' TO 'SEXTORTION:' A NUANCED INVESTIGATION OF ADOLESCENTS' ONLINE SEXUAL RISK EXPERIENCES

In prior research, the online sexual risk experiences of youth have mainly been studied in direct relation to risky sexual behaviors and coercion [9, 26, 47]. In fact, a significant line of research suggests that youth online sexual experiences can entail the most harmful outcomes including mental health, teen pregnancy, sexually transmitted diseases, and drug and alcohol misuse [9, 16, 44]. This generalized view of youths' online sexual experiences may impede risk interventions and prevention plans. For example, machine learning algorithms to detect youths' online sexual experiences could cause unintentional harm if these models were not contextualized [19]. Prior research has found that youths' online sexual experiences can be significantly vary based on the consensus state of the interaction (i.e., consensual or non-consensual) [45] and on the relationship types between the youth and the other person involved (e.g., intimate partners or strangers) [14]. Therefore, the focus of this study is to examine the role of the context on youths' online sexual experiences to inform nuanced insights into the multifaceted nature of these experiences.

2.1 Methods

In this study [6], a licensed dataset of public disclosures (N = 45, 955) made by adolescents (ages 13–17) on an online peer support platform was used to examine their online sexual risk experiences. Deep learning classifiers were trained to identify disclosures (manually labeled) (N = 8, 271) for online sexual risk experiences, and then these classifiers were contextualized to identify the levels of consent (i.e., consensual, non-consensual, sexual abuse) and relationship types (i.e., stranger, dating/friend, family). Convolutional Neural Network (CNN) performed the best for the three tasks, with an average accuracy of (AUC = 0.90). The CNN was then used to machine label the whole dataset, which resulted in identifying a total of (N = 25, 808) posts containing online sexual risk disclosures. Between-group analysis χ^2 was used to identify statistically significant differences in the proportions of the disclosures based on the levels of consent and relationship types. Then, topic modeling was applied to the disclosures divided by the two contexts (i.e., consent and relationship types). The study used the Dirichlet Mixture Model (DMM), which is specifically designed for overcoming the sparse and high-dimensional problem of clustering short texts [49]. The topic modeling results were packed by content analysis [21] to further unpack the nuances in how these online sexual risk experiences were contextualized.

2.2 Results

Youths' disclosures of sexual experiences were significantly different based on the levels of consent and relationship types. Youth were 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. Youth found to be consented participants within their online sexual experiences with their friends and dating partners, which implies a narrative shift in youth sexual education and intervention efforts to consider these interactions as normative with safety in mind. Another key finding of this study was that while youth explicitly consented to the sexual interactions with strangers, the emergent topics and the content analysis showed that mental illness indicators (e.g., depression, low self-esteem, self-harming behaviors, suicidal ideation) undercut these experiences. This finding uncovered a hidden complexity behind youth online sexual experiences that shed light on the importance of studying these experiences in tandem with other risks such as mental illness. Therefore, a broader lens of risks was adopted in my second study to incorporate mental health indicators and other offline risk behaviors to comprehensively understand modern-day youth online experiences.

3 SECOND STUDY (RQ2): PROFILING THE OFFLINE AND ONLINE RISK EXPERIENCES OF YOUTH TO DEVELOP TARGETED INTERVENTIONS FOR ONLINE SAFETY

Prior scholars have studied a wide array of youths' online and offline risk behaviors, ranging from online sexual predation to self-harm [38]. A line of research has previously investigated the impact of social media usage on youths' risk behaviors [12, 13] and mental health [25, 31]. However, many of these works have focused mainly on the impact of youths' exposure to risky content on such platforms or focused on specific risk types without providing a holistic understanding of how different risk experiences may influence one another across online and offline contexts [3, 10, 15, 23, 35]. In addition, a plethora of prior works on youth online safety has mainly relied on the self-reports of youth to examine their risk experiences, which has been presented as a limitation in this literature due to the possibility of recall bias [35, 43]. Therefore, in study two, the focus is to improve our understanding of youths' online risk experiences across online and offline contexts by aligning their self-reports with explicit risk flagging, the evidence of which can help provide youth more agency for their online/offline safety.

3.1 Methods

Under an NSF-funded project, a study was conducted called Instagram Data Donation (IGDD) [37], which was published as a case study at CHI 2022. This user study was conducted to first collect self-reports of youth (ages 13-21) using pre-validated psychological constructs to measure the frequency in which youth reported Risky

Behavior Questionnaire [8], Inventory of Statements About Self-harm [29], Cyber-Aggression Victimization [42], Cyber-Aggression Perpetration [42], Unwanted Sexual Solicitations and Approaches [33], and Youth Produced Sexual Images (Sexting) [33]. Then, the participants were asked to upload their social media (Instagram) data to self-assess the risks in their private conversations based on risk types including harassment, sexual messages or solicitations/nudity, hate speech/threat of violence, sale or promotion of illegal activities, digital self-injury, or spam. These were derived from Instagram reporting feature risk categories [1]. The participants were also asked to assess the severity levels of the risks (i.e., high, medium, and low), which were adopted based on prior literature [46]. Instagram was found to be one of the most popular and used platforms by the youth, therefore, it was the selected platform for this study [7]. In addition, the most prevalent risks that youth encounter online was found to occur within private spaces [32]; therefore, in this study, the study was conducted to collect youths' private Instagram conversations.

During the second study, 173 youth participants were able to successfully complete both parts of the study. Demographically, two thirds of participants were females (67%), (23%) were males, and (10%) were non-binary. The participants were between the ages of 13 and 21 (avg.= 17 yrs, Std.=2.15). For this study, profiles of youth (N = 173) were created based on their self-reported online and offline risk experiences using Mixture Factor Analysis (MFA) [34]. Then, between-group analysis (χ^2) was conducted to identify significant differences between the youth risk profiles based on their flagged risk types and levels. The linguistic differences in the profiles' unsafe conversations were identified using Sparse Additive Generative Model (SAGE) [17] followed by content analysis, which was used to further unpack the resulted keywords from SAGE [21].

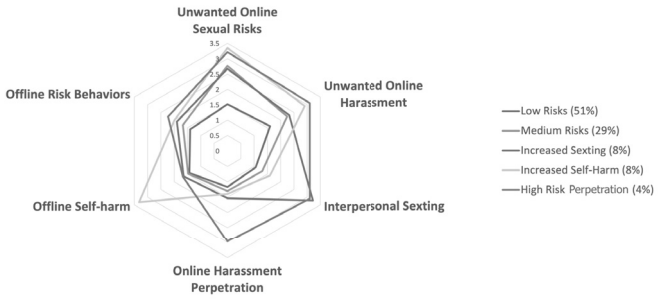

Figure 1: Youth risk profiles (N=173).

3.2 Results

Five unique profiles of youth were identified: 1) Low Risks (51% of the participants), 2) Medium Risks (29%), 3) Increased Sexting (8%), 4) Increased Self-Harm (8%), and 5) High-Risk Perpetration (4%), as shown in Figure 1. Overall, youth self-reports of online and offline risk experiences were fairly aligned with their social media trace data. Low Risks profiles who self-reported the least number of online and offline risk experiences, were found to be exposed to and flagged a wider variety of spam and scam messages and the self-harm disclosures of others. Medium Risks profile self-reported medium levels of risk experiences, yet, this profile mostly encountered harassment in the form of sexual flirtations and harassing comments. The rest of the profiles faced higher levels of risk experiences. Youth in the Increased Sexting profile self-reported the most sexting and often flagged in-person meeting requests they received within their private sexual conversations. Youth in the Increased Self-Harm profile reported the highest levels of offline self-harm, but their unsafe conversations did not contain digital self-harm content; instead, they engaged in more unsafe sexual conversations, including sugar daddy relationships with strangers. Youth that belonged to the High Risks Perpetration profile reported the highest average scores of online harassment perpetration and offline risk behaviors, including conflicts they initiated offline that escalated online, as well as promotions of illegal products. The alignment of the self-reports from youth with their social media trace data showed ecological validity. Therefore, the limitations regarding the reliance on the self-reported data in prior research may not be generalizable across all possible contexts. This study also demonstrated the importance of calibrating self-reported responses with social media data to uncover such nuances. Therefore, self-reports of online and offline risk experiences of youth will be used as features to examine their role in improving the risk detection algorithms.

4 THIRD STUDY (RQ3): REAL-TIME ONLINE RISK DETECTION FOR YOUTH GROUNDED IN CONTEXTUALIZED RISKS

Prior research on Perverted Justice dataset [2] has not been successfully able to predict sexual predatory conversations after that risk occur [11, 19], yet most of these works have not been able to detect predatory lines within the risky conversation [22]. While the models that utilize the whole conversation could capture the whole context of the conversation, these models could not protect the youth from being victimized on time [40]. On the other hand, risk detection models that rely on single messages to provide risk detection might fail to capture the context of the risky messages within the conversations [36]. As such, these models could not provide accurate and efficient risk detection within youths' private conversations in real-time. Therefore, the focus of the third study is to leverage a human-centered approach to improve real-time risk detection models for youth by using the insights from my first and second studies and creating an algorithmic contextualized framework to identify the optimal level of context that can improve detection accuracy.

4.1 Methods

For this study, the ecologically valid dataset (IGDD) [37] will be also used to train the risk detection models using youths' Instagram private conversations as well as their risk self-assessments as ground truth. From 198 participants, (N = 13, 465) conversations were flagged as safe and (N = 2, 623) were flagged as unsafe. Within

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

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

the unsafe conversations, ($N = 22,354$) messages were flagged as safe and ($N = 3,667$) were flagged as risky (including the risk types that were mentioned in study two).

The risk detection task will be performed to compare three different levels of context: 1) considering single messages only as input, 2) considering the linguistic and temporal context of the message, which could be either prior, succeeding, or both prior and succeeding messages of the risky messages based on the availability within the conversations, and 3) considering the societal and psychological contexts from youths' self-reported demographics (i.e., age, relationship status, sexual orientation, and sex) and offline and online risk experiences (stated in the second study).

Different types of Long Short-Term Memory (LSTM) networks will be applied, which can learn long-term dependencies [20], to compare the LSTMs performances for the three levels of context. LSTM shows efficient performance on Natural Language Inference (NLI) tasks such as Recognizing Textual Entailment, which identifies the relationship between two inputs [41]. Multiple LSTMs will be used: one model will read the risky message and one or two will be used to read prior and/or succeeding messages. In addition, sentence-level attention-based LSTM [48] models will be used to compare its performance with the original LSTM as well as to use the attention wights to identify what part of the messages contributes to the accuracy of the models. Lastly, conditional LSTMs [41] will be used to examine if it will improve the accuracy of the models since this type of model is conditioned based on the prior messages of the risky message. The models' accuracy results will be followed by the error analysis on the misclassified messages.

5 EXPECTED CONTRIBUTIONS AND DOCTORAL CONSORTIUM AT CHI 2023

My knowledge background is mainly in data science, so I am relatively new to the field of HCI. While applying the machine learning models for study 3 is straightforward, I am still in the stage of identifying the appropriate algorithmic contextualized framework. Therefore, attending CHI 2023 Doctoral Consortium (DC) would greatly benefit me to connect with the HCI researchers' experts to ensure that the real-time risk detection approaches for study 3 are ecologically valid and human-centered. Their feedback will be significantly helpful to strengthen the algorithmic contextual framework for these approaches. Participating in the CHI DC will be valuable for me as it will be the first doctoral consortium that I will attend. I will be honored to attend virtually to connect with other Ph.D. students from different fields to share our personal and academic experiences and challenges as Ph.D. students and discuss possible future collaborations.

Attending CHI DC will allow me to be beneficial to others as I have a strong background in Data Science, mainly related to statistical analysis and implementing machine learning algorithms. I have great experience in implementing such approaches, publishing, and reviewing works, which allows me to give feedback and discuss related topics with other participants in the DC. I will be mainly helpful in addressing the challenges when conducting ethical research or studies with youth as well as the related IRB concerns and suggesting suitable statistical measures and machine learning approaches.

ACKNOWLEDGMENTS

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

REFERENCES

[1] Zainab Agha, Reza Ghaiumy Anaraky, Karla Badillo-Urquiola, Bridget McHugh, and Pamela Wisniewski. 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*. Springer, 261–280.

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

[3] Dana Aizenkot. 2020. Social networking and online self-disclosure as predictors of cyberbullying victimization among children and youth. *Children and Youth Services Review* 119 (2020), 105695.

[4] Shiza Ali, Afsaneh Razi, Seunghyun Kim, Ashwaq Alsoubai, Joshua Gracie, Munmun De Choudhury, Pamela J Wisniewski, and Gianluca Stringhini. 2022. Understanding the Digital Lives of Youth: Analyzing Media Shared within Safe Versus Unsafe Private Conversations on Instagram. In *CHI Conference on Human Factors in Computing Systems*. 1–14.

[5] Ashwaq Alsoubai, Xavier Caddle, Ryan Doherty, Alexandra Koehler, Estefania Sanchez, Munmun De Choudhury, and Pamela Wisniewski. 2022. MOSafely, Is that Sus? A Youth-Centric Online Risk Assessment Dashboard. *CSCW'22 Companion* (2022).

[6] Ashwaq Alsoubai, Jihye Song, Afsaneh Razi, Nurun Naher, Munmun De Choudhury, and Pamela J. Wisniewski. 2021. From 'Friends with Benefits' to 'Sextortion:' A Nuanced Investigation of Adolescents' Online Sexual Risk Experiences. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2 (nov 2021). https://doi.org/10.1145/3555136

[7] Monica Anderson, Jingjing Jiang, et al. 2018. Teens, social media & technology 2018. *Pew Research Center* 31, 2018 (2018), 1673–1689.

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

[9] Aboluwaji D Ayinmoro, Endurance Uzobo, Bodisere J Teibowei, and Joyce B Fred. 2020. Sexting and other risky sexual behaviour among female students in a Nigerian academic institution. *Journal of Taibah University Medical Sciences* 15, 2 (2020), 116–121.

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

[11] Patrick Bours and Halvor Kulsrud. 2019. Detection of Cyber Grooming in Online Conversation. In *2019 IEEE International Workshop on Information Forensics and Security (WIFS)*. 1–6. https://doi.org/10.1109/WIFS47025.2019.9035090 ISSN: 2157-4774.

[12] Dawn Beverley Branley and Judith Covey. 2017. Is exposure to online content depicting risky behavior related to viewers' own risky behavior offline? *Computers in Human Behavior* 75 (2017), 283–287.

[13] Dawn Beverley Branley and Judith Covey. 2018. Risky behavior via social media: The role of reasoned and social reactive pathways. *Computers in human behavior* 78 (2018), 183–191.

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

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

[16] Pooja Chaudhary, Melissa Peskin, Jeff R Temple, Robert C Addy, Elizabeth Baumler, and Shegog Ross. 2017. Sexting and mental health: a school-based longitudinal study among youth in Texas. *Journal of Applied Research on Children* 8, 1 (2017), 11.

[17] Jacob Eisenstein, Amr Ahmed, and Eric P Xing. 2011. Sparse additive generative models of text. In *Proceedings of the 28th international conference on machine learning (ICML-11)*. 1041–1048.

[18] Vaishali U Gongane, Mousami V Munot, and Alwin D Anuse. 2022. Detection and moderation of detrimental content on social media platforms: current status and future directions. *Social Network Analysis and Mining* 12, 1 (2022), 1–41.

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

[20] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[21] Hsiu-Fang Hsieh and Sarah E Shannon. 2005. Three approaches to qualitative content analysis. *Qualitative health research* 15, 9 (2005), 1277–1288.

[22] Giacomo Inches and Fabio Crestani. 2012. Overview of the International Sexual Predator Identification Competition at PAN-2012.. In *CLEF (Online working notes/labs/workshop)*, Vol. 30.

[23] Dylan B Jackson, Cashen M Boccio, and Wanda E Leal. 2020. Do youth who vape exhibit risky health lifestyles? Monitoring the future, 2017. *Preventive medicine* 136 (2020), 106101.

[24] Alejandro Jaimes, Daniel Gatica-Perez, Nicu Sebe, and Thomas S Huang. 2007. Guest Editors' Introduction: Human-Centered Computing–Toward a Human Revolution. *Computer* 40, 5 (2007), 30–34.

[25] Betul Keles, Niall McCrae, and Annmarie Grealish. 2020. A systematic review: the influence of social media on depression, anxiety and psychological distress in adolescents. *International Journal of Adolescence and Youth* 25, 1 (2020), 79–93.

[26] Poco D Kernsmith, Bryan G Victor, and Joanne P Smith-Darden. 2018. Online, offline, and over the line: Coercive sexting among adolescent dating partners. *Youth & Society* 50, 7 (2018), 891–904.

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

[28] Seunghyun Kim, Afsaneh Razi, Gianluca Stringhini, Pamela J Wisniewski, and Munmun De Choudhury. 2021. You Don't Know How I Feel: Insider-Outsider Perspective Gaps in Cyberbullying Risk Detection.. In *ICWSM*. 290–302.

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

[30] Sonia Livingstone and Ellen Helsper. 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 (March 2010), 309–329. https://doi.org/10.1177/1461444809342697

[31] Francesco Lupariello, Serena Maria Curti, Elena Coppo, Sara Simona Racalbuto, and Giancarlo Di Vella. 2019. Self-harm risk among adolescents and the phenomenon of the "Blue Whale Challenge": case series and review of the literature. *Journal of forensic sciences* 64, 2 (2019), 638–642.

[32] Kimberly J Mitchell, David Finkelhor, and Janis Wolak. 2007. Youth Internet users at risk for the most serious online sexual solicitations. *American Journal of Preventive Medicine* 32, 6 (2007), 532–537.

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

[34] Bengt Muthén and Bengt O Muthén. 2009. *Statistical analysis with latent variables*. Wiley New York, NY.

[35] Anthony T Pinter, Pamela J Wisniewski, Heng Xu, Mary Beth Rosson, and Jack M Caroll. 2017. Adolescent online safety: Moving beyond formative evaluations to designing solutions for the future. In *Proceedings of the 2017 Conference on Interaction Design and Children*. 352–357.

[36] Afsaneh Razi, Ashwaq Alsoubai, Seunghyun Kim, Shiza Ali, Gianluca Stringhini, Munmun De Choudhury, and Pamela J. Wisniewski. 2023. Sliding into My DMs: Detecting Uncomfortable or Unsafe Sexual Risk Experiences within Instagram Direct Messages Grounded in the Perspective of Youth. *Proc. ACM Hum.-Comput. Interact.* CSCW2 (Accepted, but Not Published 2023).

[37] Afsaneh Razi, Ashwaq AlSoubai, Seunghyun Kim, Nurun Naher, Shiza Ali, Gianluca Stringhini, Munmun De Choudhury, and Pamela J Wisniewski. 2022. 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*. 1–9.

[38] Afsaneh Razi, Karla Badillo-Urquiola, and Pamela J Wisniewski. 2020. Let's Talk about Sext: How Adolescents Seek Support and Advice about Their Online Sexual Experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[39] Afsaneh Razi, Seunghyun Kim, Ashwaq Alsoubai, Xavier Caddle, Shiza Ali, Gianluca Stringhini, Munmun De Choudhury, and Pamela Wisniewski. 2021. 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*.

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

[41] Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomáš Kočiskỳ, and Phil Blunsom. 2015. Reasoning about entailment with neural attention. *arXiv preprint arXiv:1509.06664* (2015).

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

[43] Michael A Tarrant, Michael J Manfredo, Peter B Bayley, and Richard Hess. 1993. Effects of recall bias and nonresponse bias on self-report estimates of angling participation. *North American Journal of Fisheries Management* 13, 2 (1993), 217–222.

[44] Joris Van Ouytsel, Michel Walrave, Lieven De Marez, Bart Vanhaelewyn, and Koen Ponnet. 2020. A first investigation into gender minority adolescents' sexting experiences. *Journal of Adolescence* 84 (2020), 213–218.

[45] Sebastian Wachs, Michelle F Wright, Manuel Gámez-Guadix, and Nicola Döring. 2021. How are consensual, non-consensual, and pressured sexting linked to depression and self-harm? The moderating effects of demographic variables. *Int'l journal of environmental research and public health* 18, 5 (2021), 2597.

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

[47] Janis Wolak, David Finkelhor, Wendy Walsh, and Leah Treitman. 2018. Sextortion of minors: Characteristics and dynamics. *Journal of Adolescent Health* 62, 1 (2018), 72–79.

[48] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*. 1480–1489.

[49] Jianhua Yin and Jianyong Wang. 2014. A dirichlet multinomial mixture model-based approach for short text clustering. *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining* (2014).