Deploying Human-Centered Machine Learning to Improve Adolescent Online Sexual Risk Detection Algorithms

Afsaneh Razi
Department of Computer Science
University of Central Florida
Orlando, FL 32826,USA
Afsaneh.razi@knights.ucf.edu

GROUP '20 Companion, January 6–8, 2020, Sanibel Island, FL, USA

© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-6767-7/20/01…$15.00
https://doi.org/10.1145/3323994.3372138

Abstract
As adolescents' engagement increases online, it becomes more essential to provide a safe environment for them. Although some apps and systems are available for keeping teens safer online, these approaches and apps do not consider the needs of parents and teens. We would like to improve adolescent online sexual risk detection algorithms. In order to do so, I'll conduct three research studies for my dissertation: 1) Qualitative analysis on teens posts on an online peer support platform about online sexual risks in order to gain deep understanding of online sexual risks 2) Train a machine learning approach to detect sexual risks based on teens conversations with sex offenders 3) develop a machine learning algorithm for detecting online sexual risks specialized for adolescents.

Author Keywords
Adolescent Online Safety; Human-centered Machine Learning; Online sexual risks.

ACM Classification Keywords
• Human-centered computing~Empirical studies in HCI

Motivation and Background

As adolescents have access to the internet more than ever, protecting and providing a safe online experience becomes more important. According to Pew Research Center [10], about half of the teens (45%) are 'almost constantly' online and most (71%) have social media account. Although the internet provides a lot of opportunities for teens, it also exposes them to online risks [6]. According to Crimes Against Children Research Center, 1 in 11 teens in the U.S. have experienced unwanted sexual solicitations, 1 in 9 are victims of cyberbullying, and 1 in 4 are exposed to explicit sexual materials online [7]. One of the most common types of online risks is online sexual risks [4]. Available solutions for adolescent online safety are mostly device-based restrictions and parental monitoring apps [3]. Though these approaches do not take into consideration the context of which risks happen, so they overwhelm parents with huge amount of teens' data. Also they are privacy-invasive to teens [3]. Therefore, risk detection algorithms need to take into account the context that an online risk happens to be more useful for parents and less privacy-invasive to teens.

Machine learning algorithms need high-quality training datasets to learn key attributes of data in order to create intelligent systems to detect a particular phenomenon of interest [9]. So, the ground truth or labels of a training dataset determines how much a risk detection algorithm is effective. Most of the training datasets used in prior research to detect online sexual risks are not representative of or generalizable to adolescents because there are not based on teens' real-world data.

In addition, most of the qualitative research on the topic of adolescent online safety has historically relied on teen-self-reports than utilizing social media trace data [8]. Although usually in these self-reported studies participants were assured that their answers would remain anonymous, it might be possible that teens do not want to disclose this information and provide socially desirable answers. Therefore, utilizing real-world adolescent social media data to analyze sexual behaviors will help the community to have a better understanding of adolescent sexual risks.

In this dissertation study, we direct our attention toward a better understanding of the sexual content that adolescents exchange online and developing algorithms that can detect these risks accurately. The primary goals of this research are to 1) develop a better understanding of how adolescents engage in online sexual experiences. 2) Create contextual training datasets for adolescent online risk detection. 3) Improve online risk detection systems.

The following are associated research questions that we are going to examine:

RQ1- *What types of sexual experiences and risks do teens experience online?*

RQ2- *How qualitative insight could be operationalized into viable sexual risk models for labeling datasets in which to train algorithms to detect adolescent online sexual risks?*

RQ3- *How can adolescent online sexual risk detection algorithms be improved?*

Proposed Methods

In my dissertation study, I am planning to use different types of methods in order to answer the research questions and tackle the problem. This includes using qualitative analysis in order to understand online sexual risks better, and then create rich training dataset to train machine learning models on that.

First study (Addressing RQ1): For my first dissertation study, I conducted a thematic content analysis of 4,180 posts by teens ages 12-17 on an online peer support forum to uncover what and how teens talk about their online sexual interactions with others. We found that teens used the platform to seek support (83%), connect with others (15%), and give advice (5%) about sexting, their sexual orientation, sexual abuse, and explicit content. In addition, we found that female teens often received unwanted nudes from strangers. But their main struggle was with how to turn down sexting requests from people they knew such as friends and their significant other. I conducted this research to gain more insight into the online sexual experiences of teens and how they seek support around these issues. By understanding the online sexual experiences of teens, we can create richer training sets based on their real experiences.

Second study (Addressing RQ2 and RQ3): Most of the research in the area of online sexual risk detection [1,2,5] used perverted justice [11] public dataset from convicted sex offender and volunteers acting as teens. Almost all of the research in this area focused on detecting the predators with acceptable accuracy of 93% form PAN competition [12]. Though researchers where not successful on identifying predatory lines (47% accuracy). My aim is to examine predatory conversations from victims' perspectives. Also, I would want to improve the algorithm for finding predatory lines. In order to do this, I will create a rich dataset to serve our purpose then I will develop appropriate machine learning to detect online predatory risks.

Third study (Addressing RQ2 and RQ3): In the last study, my objective is to collect Instagram data of teens in order to train and evaluate risk detection algorithms. Analyzing Instagram data is chosen because it is used by more than 70% of adolescents in the U.S.[10]. Instagram and YouTube are the top social media platforms being used by half of U.S. teens ages 13 to 17 [10]. The aim is to create a robust training dataset using the teen social media data. Then to establish ground truth labels on riskiness of social media utilizing the previous qualitative data coding. In this study first we have them fill an online survey about their social media use, online risk experience, and their mental health. Then we ask them to upload their Instagram data on our website. Then we show them their conversations on Instagram and ask them to choose messages that made them feel uncomfortable or unsafe. We designed these surveys so we can have a ground truth of what teens think it is risky and associate that with their real social media data.

Research Status

My first research study for my dissertation have been accepted to the ACM CHI Conference on Human Factors in Computing Systems 2020. For the second study I am planning to create the training dataset by summer 2020 from Perverted Justice dataset and apply machine learning algorithm. For the third study we have developed a website for the surveys and the Instagram data collection which we lunch on Spring 2020. I will

create a training dataset by Fall 2020 and develop accurate machine learning sexual risk detection algorithms by Spring 2020.

Expected Contributions and Doctoral Colloquium at GROUP 2020

The novel contribution of this thesis is to leverage machine learning techniques and human-centered approaches for the design and development of tools that address the problem of adolescent exposure to online sexual risks. I will use a deeper contextual understanding of teen social media users by leveraging human insights to develop online sexual risk detection algorithms tailored to adolescents. I expect that my approach will be more helpful for teens and their parents. It would be valuable for me to get feedback on how to strengthen my dissertation from perspective of ideas and methods for tackling the problem. Attending the DC at GROUP would be beneficial for me in terms of learning more methodologies and practices from stablished researchers in SIGCHI community.

Acknowledgement

This research is supported by the U.S. National Science Foundation under grant number #IIP-1827700 and by the William T. Grant Foundation under grant number #187941. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the research sponsors.

References

1. Y. Cheong, A. K. Jensen, E. R. Guðnadóttir, B. Bae, and J. Togelius. 2015. Detecting Predatory Behavior in Game Chats. *IEEE Transactions on Computational Intelligence and AI in Games* 7, 3: 220–232.

2. Mohammadreza Ebrahimi, Ching Y. Suen, and Olga Ormandjieva. 2016. Detecting predatory conversations in social media by deep Convolutional Neural Networks. *Digital Investigation* 18: 33–49.

3. Arup Kumar Ghosh, Karla Badillo-Urquiola, Shion Guha, Joseph J. LaViola Jr, and Pamela J. Wisniewski. 2018. Safety vs. Surveillance: What Children Have to Say about Mobile Apps for Parental Control. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*, ACM Press, 1–14.

4. Juliane A. Kloess, Anthony R. Beech, and Leigh Harkins. 2014. Online child sexual exploitation: prevalence, process, and offender characteristics. *Trauma, Violence & Abuse* 15, 2: 126–139.

5. Carlos Laorden, Patxi Galán-García, Igor Santos, Borja Sanz, Jose Gomez Hidalgo, and Pablo Bringas. 2013. Negobot: A Conversational Agent Based on Game Theory for the Detection of Paedophile Behaviour. In *Advances in Intelligent Systems and Computing*. 261–270.

6. Sonia Livingstone and Peter K. Smith. 2014. Annual Research Review: Harms experienced by child users of online and mobile technologies: the nature, prevalence and management of sexual and aggressive risks in the digital age. *Journal of Child Psychology and Psychiatry* 55, 6: 635–654.

7. Kimberly Mitchell, Lisa Jones, David Finkelhor, and Janis Wolak. 2014. Trends in Unwanted Online Experiences and Sexting : Final Report. *Crimes Against Children Research Center*.

8. 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. *Proceedings of the 2017 Conference on Interaction Design and Children*, ACM, 352–357.

9. Mattias Rost, Louise Barkhuus, Henriette Cramer, and Barry Brown. 2013. Representation and communication: challenges in interpreting large social media datasets. 6.

10. 2018. Teens, Social Media & Technology 2018 | Pew Research Center. Retrieved September 22, 2018 from http://www.pewinternet.org/2018/05/31/teens-social-media-technology-2018/.

11. Perverted-Justice.com - The largest and best anti-predator organization online. Retrieved October 28, 2019 from http://www.perverted-justice.com/.

12. PAN at CLEF 2012 - Author Identification. Retrieved November 1, 2019 from https://pan.webis.de/clef12/pan12-web/author-identification.html.