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

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

Xavier Caddle
xavier.caddle@knights.ucf.edu
University of Central Florida
Orlando, U.S.A

Ryan Doherty
rdoherty20@knights.ucf.edu
University of Central Florida
Orlando, Florida, USA

Alexandra Koehler
chibitalex@knights.ucf.edu
University of Central Florida
Orlando, Florida, USA

Estefania Sanchez
estefania.sanchez@knights.ucf.edu
University of Central Florida
Orlando, Florida, USA

Munmun De Chodhury
munmund@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, U.S.A

Pamela J. Wisniewski
pamela.wisniewski@vanderbilt.edu
Vanderbilt University
Nashville, Tennessee, USA

ABSTRACT

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.

CCS CONCEPTS

• **Online Safety** → **Artificial Intelligence** ; • **Youth-Centered Artificial Intelligence** → *Risk Detection Dashboard*;

ACM Reference Format:
Ashwaq Alsoubai, Xavier Caddle, Ryan Doherty, Alexandra Koehler, Estefania Sanchez, Munmun De Chodhury, and Pamela J. Wisniewski. 2022. MOSafely, *Is that Sus?* A Youth-Centric Online Risk Assessment Dashboard. In *Companion Computer Supported Cooperative Work and Social Computing (CSCW'22 Companion), November 8–22, 2022, Virtual Event, Taiwan.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3500868.3559710

CSCW'22 Companion, November 8–22, 2022, Virtual Event, Taiwan

ACM ISBN 978-1-4503-9190-0/22/11.
https://doi.org/10.1145/3500868.3559710

1 INTRODUCTION

Adolescents' social growth and developmental exploration are mainly mediated through extensive social media usage [4]. Although social media provides youth a unique opportunity to communicate and learn, it also exposes them to a wide array of risks that could have adverse consequences [10]. A major trend of the current approaches for adolescents' online safety is relying on parental control that are not only privacy-invasive to youth, but also overload parents with unnecessary information [1, 6]. Today, Artificial Intelligence (AI)-based risk detection technologies present promising potentials to automatically detect risky content [20]. However, these models could pose a digital inequity especially for socioeconomically disadvantaged youth [15]. Thus, human-computer interaction (HCI) researchers have been advocating for building AI online safety solutions that are youth-centric [15]. 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.

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 [7]. 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 [17]. Furthermore, the majority of the presented risk detection approaches in youth online safety literature lack the youth engagement of these approaches [9, 16]. 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 [16]. 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 [5]. 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).

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.

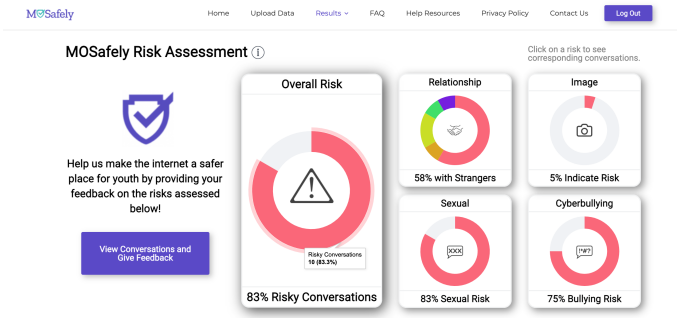

Figure 1: Screenshot of the risk assessment dashboard after users successfully uploaded their social media file.

- **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 [19]. As such, we used systematic reviews of automatic (machine learning based) detection approaches for these risks [9, 16] 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 [3], the relationship type classifier was designed to only 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 [16]. 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 [11] 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.

4 TECHNICAL IMPLEMENTATION

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

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,

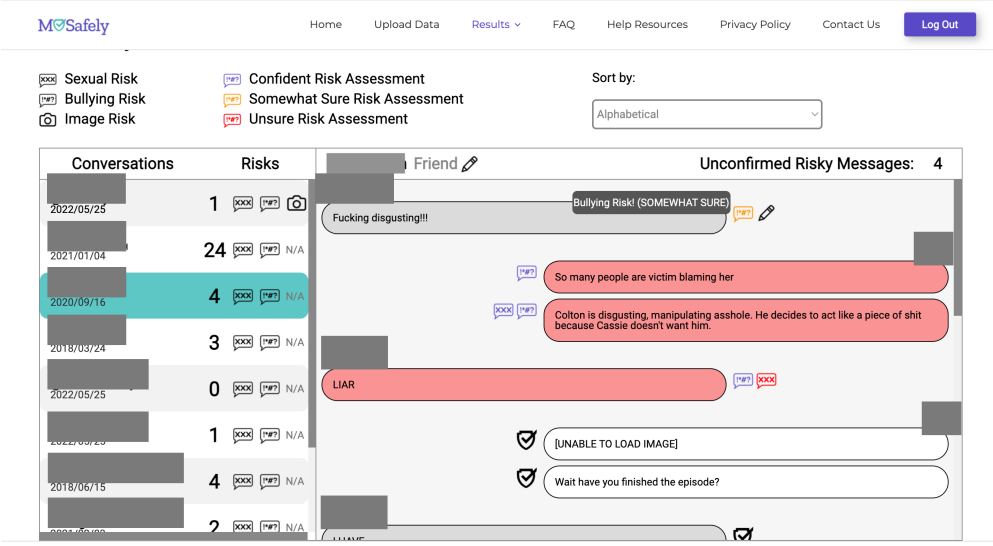

Figure 2: Screenshot of the conversation page, showing the edit icon for the message level feedback.

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 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

we collected an ecologically valid dataset that consisted of youth private conversations along with their risk flags to their conversations; [14] 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 [2, 8, 13], 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 1. The best performing models for each risk type were then compiled as TensorFlow saved 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.

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.

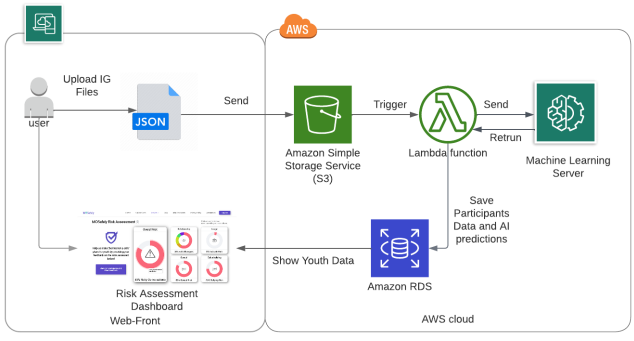

Figure 3: MOSafely Architecture.

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

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 [18]. 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 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.

ACKNOWLEDGMENTS

We thank Afsaneh Razi, Seunghyun Kim, and Shiza Ali who provided the skeleton algorithms for the dashboard and Kenneth Tran and Zach Arehart who assisted with the development of this system. 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] Mamtaj Akter, Amy J Godfrey, Jess Kropczynski, Heather R Lipford, and Pamela J Wisniewski. 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 (2022), 1–28.

[2] 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.

[3] 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.

[4] Monica Anderson and Jingjing Jiang. 2018. Teens, Social Media & Technology 2018 | Pew Research Center. http://www.pewinternet.org/2018/05/31/teens-social-media-technology-2018/

[5] Stevie Chancellor, Eric PS Baumer, and Munmun De Choudhury. 2019. 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 (2019), 1–32.

[6] Arup Kumar Ghosh, Karla Badillo-Urquiola, Mary Beth Rosson, Heng Xu, John Carroll, and Pamela J. Wisniewski. 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. https://doi.org/10.1145/3173574.3173768

[7] Haiyan Jia, Pamela J Wisniewski, Heng Xu, Mary Beth Rosson, and John M Carroll. 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*. 583–599.

[8] Seunghyun Kim, Afsaneh Razi, Ashwaq Alsoubai, Chen Ling, Gianluca Stringhini, Pamela J. Wisniewski, and Munmun De Choudhury. 2022. I'm Talking to You: Detecting and Differentiating between Online Harassment in Networked Public vs. Private Social Media Spaces.

[9] 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

[10] 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

[11] Robert Munro Monarch. 2021. *Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI*. Simon and Schuster.

[12] Rachel Nosowsky and Thomas J Giordano. 2006. The Health Insurance Portability and Accountability Act of 1996 (HIPAA) privacy rule: implications for clinical research. *Annu. Rev. Med.* 57 (2006), 575–590.

[13] Afsaneh Razi, Ashwaq AlSoubai, Seunghyun Kim, Shiza Ali, Gianluca Stringhini, Munmun De Choudhury, and Pamela J Wisniewski. 2022. Sliding into My DMs: Detecting Uncomfortable or Unsafe Sexual Risk Experiences within Instagram Direct Messages Grounded in the Perspective of Youth.

[14] 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.

[15] Afsaneh Razi, Seunghyun Kim, Ashwaq Alsoubai, Xavier Caddle, Shiza Ali, Munmun De Choudhury, Pamela Wisniewski, et al. 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*.

[16] 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

[17] Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller. 2019. *Explainable AI: interpreting, explaining and visualizing deep learning*. Vol. 11700. Springer Nature.

[18] Ben Shneiderman, Catherine Plaisant, Maxine S Cohen, Steven Jacobs, Niklas Elmqvist, and Nicholas Diakopoulos. 2016. *Designing the user interface: strategies for effective human-computer interaction*. Pearson.

[19] 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*. 3919–3930.

[20] Agnieszka Wrońska, Rafał Lew-Starowicz, and Anna Rywczyńska. 2020. *Education–Relationships–Play Multifaceted Aspects of the Internet and Child and Youth Online Safety*. Foundation for the Development of the Education System.

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