Sliding into My DMs: Detecting Uncomfortable or Unsafe Sexual Risk Experiences within Instagram Direct Messages Grounded in the Perspective of Youth

AFSANEH RAZI, Drexel University, U.S.A
ASHWAQ ALSOUBAI*, Vanderbilt University, U.S.A
SEUNGHYUN KIM*, Georgia Institute of Technology, U.S.A
SHIZA ALI, Boston University, U.S.A
GIANLUCA STRINGHINI, Boston University, U.S.A
MUNMUN DE CHOUDHURY, Georgia Institute of Technology, U.S.A
PAMELA J. WISNIEWSKI, Vanderbilt University, U.S.A

We collected Instagram data from 150 adolescents (ages 13-21) that included 15,547 private message conversations of which 326 conversations were flagged as sexually risky by participants. Based on this data, we leveraged a human-centered machine learning approach to create sexual risk detection classifiers for youth social media conversations. Our Convolutional Neural Network (CNN) and Random Forest models outperformed in identifying sexual risks at the conversation-level (AUC=0.88), and CNN outperformed at the message-level (AUC=0.85). We also trained classifiers to detect the severity risk level (i.e., safe, low, medium-high) of a given message with CNN outperforming other models (AUC=0.88). A feature analysis yielded deeper insights into patterns found within sexually safe versus unsafe conversations. We found that contextual features (e.g., age, gender, and relationship type) and Linguistic Inquiry and Word Count (LIWC) contributed the most for accurately detecting sexual conversations that made youth feel uncomfortable or unsafe. Our analysis provides insights into the important factors and contextual features that enhance automated detection of sexual risks within youths' private conversations. As such, we make valuable contributions to the computational risk detection and adolescent online safety literature through our human-centered approach of collecting and ground truth coding private social media conversations of youth for the purpose of risk classification.
Content Warning: *This paper discusses sensitive topics, such as sex, which may be triggering.*

CCS Concepts: • **Human-centered computing** → **Empirical studies in HCI**; **Social networking sites**; **Empirical studies in collaborative and social computing**; • **Computing methodologies** → **Machine learning algorithms**; *Classification and regression trees*; *Neural networks*; *Natural language processing*; • **Security and privacy** → **Social aspects of security and privacy**.

Additional Key Words and Phrases: Sexual Risk Detection, Adolescents Online Safety, Youth Online Risks, Machine Learning, Deep Learning

*Both authors contributed equally to this research.

Authors' addresses: Afsaneh Razi, afsaneh.razi@drexel.edu, Drexel University, 3675 Market St 10th floor, Philadelphia, Pennsylvania, U.S.A, 19104; Ashwaq AlSoubai, Vanderbilt University, 2201 West End Ave, Nashville, TN, U.S.A, ashwaq.alsoubai@vanderbilt.edu; Seunghyun Kim, Georgia Institute of Technology, 30318, Atlanta, Georgia, U.S.A, seunghyun.kim@gatech.edu; Shiza Ali, Boston University, 02215, Boston, Massachusetts, U.S.A, shiza@bu.edu; Gianluca Stringhini, Boston University, 02215, Boston, Massachusetts, U.S.A, gian@bu.edu; Munmun De Choudhury, Georgia Institute of Technology, 30318, Atlanta, Georgia, U.S.A, munmund@gatech.edu; Pamela J. Wisniewski, Vanderbilt University, 2201 West End Ave, Nashville, TN, U.S.A, pamela.wisniewski@vanderbilt.edu.

2573-0142/2023/4-ART89 $15.00
https://doi.org/10.1145/3579522

ACM Reference Format:
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.* 7, CSCW1, Article 89 (April 2023), 29 pages. https://doi.org/10.1145/3579522

1 INTRODUCTION

In 2020, more than 21.7 million reports of suspected child sexual exploitation were made to the National Center for Missing and Exploited Children's CyberTipline, which increased by 97% compared to the year prior [1]. With the rise in computer-mediated sexual risks, the Human-Computer Interaction (HCI) and Artificial Intelligence (AI) research communities have collectively worked towards understanding how these sexual risks unfold and can be prevented, ranging from in-depth qualitative accounts of sexual victimization [6, 22, 23, 31, 32] to computational approaches for sexual risk detection [47, 71]. For instance, the #MeToo movement [34] gave rise to a body of work where researchers began to detect sexual harassment/abuse within public social media posts [28]. The culmination of increased sexual exploitation of youth online and the rise in state-of-the-art computational risk detection approaches for sexual exploitation produce a timely and critical opportunity to leverage the CSCW community's strengths to actively protect youth online.

A recent review of the computational approaches to sexual risk detection synthesized this growing body of literature and called for a more human-centered approach to machine learning (HCML) to move the field forward in a way that would affect real societal impact [60]. For instance, the review highlighted the need for collecting ecologically valid datasets for training robust classifiers to make accurate predictions relevant to real people and contexts. The extant research tended to focus on publicly available datasets, while the most concerning sexual risks such as sexual solicitation and harassment occur in private online spaces like instant messaging and chat rooms [80]. Further, sexual risk classifiers often did not take into account survivors' accounts of their own risk experiences; instead, they often relied heavily on third-party annotators to identify cases of sexual victimization [60]. As risk is a highly subjective construct [54], quantifiably operationalizing sexual victimization for the purpose of risk detection is difficult without direct input from the individual who experienced it. Finally, existing approaches primarily leveraged linguistic and semantic cues but rarely considered human-centered insights in terms of the contextual factors that have been shown in the literature to increase one's susceptibility to be sexually victimized or groomed [78]. In our case, relevant contextual factors for youth may include developmental (e.g., age), individual (e.g., gender), and relational (e.g., nature of the relationship) factors that have been found to be salient to increase sexual victimization in the adolescent online safety and risk literature [5, 24, 59, 78]. We posit that it is important to take youths' perspectives of their sexual risk experiences into consideration, so that we can identify contextual features important for risk detection. To do this, we analyzed Instagram Direct Messages (DM's) of youth at both the conversation-level (i.e., all messages exchanged in a given private chat) and message-level (i.e., an individual DM) to address the following research questions:

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

To answer these questions, we collected Instagram data from 150 adolescents (ages 13-21) and asked them to flag their own private messages for sexual content that made them feel uncomfortable or unsafe which included 15,547 conversations and 326 of those were flagged as containing sexual risks. We trained and tested conversation-level sexual risk classifiers (RQ1), and found Convolutional Neural Network (CNN) outperformed traditional models (accuracy=0.89). For traditional models, the Random Forest model that incorporated age, gender, and relationship type as contextual features with linguistic features outperformed other models with an accuracy of 0.88. Next, we developed a message-level classifiers for predicting whether a given message contained sexually risky content with an accuracy of 0.84, as well as the level of risk posed to the victim (i.e., safe, low, medium-high) with an accuracy of 0.82 (RQ2). To answer RQ3, we unpacked how the contextual features and psycholinguistic attributes (based on the Linguistic Inquiry and Word Count, LIWC [52]) played a role in the online sexual experiences of youth. Young adults were significantly more likely to flag conversations as safe, while young teens (between 13-15) and adolescents (16-18) flagged more unsafe sexual conversations. Safe conversations were more likely to be between the participants and family members, friends, or significant others, while unsafe conversations were significantly more likely between participants and strangers or acquaintances. Next, we analyzed the relative importance among LIWC categories and found distinguishing LIWC categories for unsafe and safe conversations. For instance, unsafe conversations contained more words from the "friends" category (e.g., friend, neighbor), compared to safe conversations that contained more words from the "family" category (e.g., sister, daughter). Additionally, an error analysis helped us identify that most of the misclassified instances were due to short conversations that included links or media. Our analysis sheds light on the salient features to leverage in sexual risk detection algorithms, as well as the online sexual risk experiences of youth. Overall, our research makes the following contributions to the Computer-Supported Cooperative Work And Social Computing (CSCW) research community:

- We took great care and effort to create an ecologically valid dataset based on private social media conversations of youth. The dataset was labeled by youth from their own perspective of sexual risks, spanning incidents that may have made them feel uncomfortable or unsafe. Importantly, this work generated a valuable dataset based on real-world situations that can be utilized in future research. It further demonstrates that it is possible to build machine learning classifiers to distinguish unsafe sexual messages and reveals their key differences.
- We went beyond identifying sexual predators or detecting sexual harassment in public posts by building classifiers to assess sexual risk in private conversations of youth. In particular, we developed automated machine learning-based detection models that could act as a key element in ensuring online safety of youth. In addition, we built machine learning (ML) approaches to predict the presence and severity of sexual risks in conversations as well as their constituent messages, followed by highlighting their differences.
- Our findings shine a light on the importance of contextual features (e.g., age, gender, and relationship type) in identifying sexually risky conversations, and how automated sexual risk detection models could utilize them for more human-centered risk detection systems.
- We suggest important design implications for computational approaches for detecting sexual risks in private conversations. Additionally, we contribute to the youth online safety by human-insights relevant to youth unsafe sexual interactions.

2 RELATED WORK

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

2.1 Computational Sexual Risk Detection Literature

The majority of computational sexual risk detection research has been conducted in the context of sexual grooming and identification of child sexual predators (75%), sex trafficking (12%), and sexual harassment and/or abuse of adults (12%) [60]. Much of this work started with utilizing traditional ML approaches during the 2012 Sexual Predator Identification competition ran by PAN [1] [37]. After the competition, researchers continued the effort by presenting different traditional models to detect child sexual predators in the PAN-12 data [26]. A relatively smaller subset of the literature adopted deep learning methods for detecting sexual harassment, abuse, or sex trafficking [60]. For instance, researchers compared the performances of deep learning models on a publicly-available dataset "SafeCity," which includes stories for sexual harassment disclosure detection [39, 47]. While several of these studies achieved high performance, most benchmarked their performance based solely on ML metrics (e.g, accuracy, F1-score, recall, precision) [60]. Although these performance metrics are important to evaluate the accuracy of the models, it is important to consider the social interpretation behind the algorithms to thoroughly evaluate the models in real use [10].

Another theme within previous research on sexual risk detection was that most researchers have mainly focused on predicting risk as a binary task (risky vs. non-risky) instead of considering different risk levels [60]. Yet, what we know about risks posed to youth online is that it is a spectrum that can escalate over time [38], rather than a dichotomous state. Thus, some researchers have tried to differentiate risk by differing levels. Ringenberg et al. [64] used Fuzzy Sets for labeling messages for three levels of risks (low, medium, high), and developed Neural Network models that used these fuzzy membership functions of each line in a chat as input to predict the risky interaction. CNN was found in this work as the best model for predicting risk levels. While Seigfried-Spellar et al. [67] classified conversations from the Perverted Justice (PJ) dataset[2] based on two risk levels for a contact offense which is determined on the model's predicted probabilities of whether the offender showed-up to meet the decoy in the physical world. Therefore, identifying the risk levels could provide in-depth information on the potential degree of the harm to the youth so proper risk mitigation strategies could be used than just a binary classification of whether the risk exists. Therefore, in our study we leveraged ML algorithms to be trained on the conversation-level to identify the unsafe sexual conversations and went beyond that to train the models on the message-evel to identify the risk levels (low, medium, and high) within these messages.

2.2 Leveraging HCML to Improve Sexual Risk Detection for Youth

As AI has become an irrevocable part of systems that influence peoples' lives, concerns about uncertainty and potential mistakes made by these systems has become heightened [76]. For instance, AI has been used to identify child predators online by developing a deep understanding of the linguistic cues used in the process of sexual grooming [15, 48]. Yet, without an evidence-based understanding of grooming behaviors, risk detection algorithms could be harmful to those classified as alleged predators (e.g., due to false positives) or potential victims (e.g., in the case of false negatives). We address these gaps by taking a Human-Centered Machine Learning (HCML) approach to detect sexual risks encountered online by youth. HCML keeps humans at the center of the design process by taking into account stakeholders' needs, as well as their perspectives. Leveraging practices from HCML [76] is needed to ensure that knowledge about people is used to create robust algorithms and consider potential mistakes/harms that these systems might make [65].

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

[1] A benchmarking activity on uncovering plagiarism, authorship and social software misuse http://pan.webis.de
[2] http://www.perverted-justice.com/

to our research. First, datasets traditionally analyzed for sexual harassment or abuse were mostly based on public posts on social media, such as Twitter [60]. The most popular public datasets used in the literature for identifying sexual groomers, for instance, utilize the PJ dataset and PAN-2012 competition dataset, which was created from PJ with combination of other datasets. The PJ dataset includes logs of online conversations between convicted sex offenders and adult volunteers posing as minors, which is not representative of real-world data from youth. Analyzing public discourse to understand sexual harassment and abuse is another problem, since it is well known that people often behave differently in public spaces than they do privately. Since most sexual risks occur in private channels [80], it is important to examine these interactions. Second, most risk detection systems have relied on labels that have not been grounded in the victim's perception of risks. Past literature relies heavily on third-party annotations [60], although their perspective of risks might be different than the actual victims. For instance, Kim et al. [40] has found that ML models for detecting cyberbullying instances trained based on the perspectives of victims of bullying (i.e., "insiders") outperformed the models trained on data annotated by third-party annotators (i.e., "outsiders") by detecting implicit references to bullying. Thus, incorporating first-person perspectives in ground truth labeling of one's sexual risk experiences is an important step towards establishing a risk detection system that does not estrange the key stakeholders of the system.

Finally, the majority of studies on online sexual risk detection have primarily focused on the textual features which represent the linguistic style embedded in the text [60]. The dominance of these textual features such as N-grams, bag-of-words (BoW), and word embeddings entail the *what* and *how* of the dataset; however, this falls short of encompassing the crucial question of *who* is involved in the specific conversation, post, or comment. Since different people perceive differently based on their life experiences [29], it is important to incorporate the human-centered features, such as age and gender of the individuals receiving the message (i.e., our participants), in the training of risk detection systems. Therefore, identifying and utilizing the social and psychological patterns as indicators of risks may be beneficial in developing more effective risk detection models.

3 STUDY DESIGN AND DATASET

We present the design of our study, including the arduous process for data collection and verification, as well as characteristics about our participants and their Instagram data.

3.1 Social Media Data Collection

We recruited participants between the ages of 13-21 who were: 1) English speakers based in the United States, 2) had an active Instagram account currently and for at least 3 months during the time they were 13-17 years-old, 3) exchanged Direct Messages (DMs) with at least 15 people, and 4) had at least 2 DMs that made them or someone else feel uncomfortable or unsafe, and 5) were willing to share their Instagram data with us for the purpose of research. If eligible, participants over the age of 18 provided their own informed consent to enroll in the study, while minors were asked to assent to participation with their parent or legal guardian's required informed consent.

Once enrolled in the study, participants were first asked to complete a web-based survey to provide their demographic information and some additional details about their online experiences. Then, they were asked to request their Instagram data file in the form of zipped JSON files to upload to our system. We selected Instagram as the social media platform of choice, since more than half (72%) of teens use Instagram, making it one of the most popular social media platforms among youth [7]. We leveraged Amazon Web Services (AWS), RDS, EC2, PHP, Python, and other technologies to develop a secure web-based data collection system and stored their DMs in a relational database to present back to them in the interface. Having participants review and annotate their own DMs was necessary for establishing ground truth for sexual risk detection, as described below. Once the

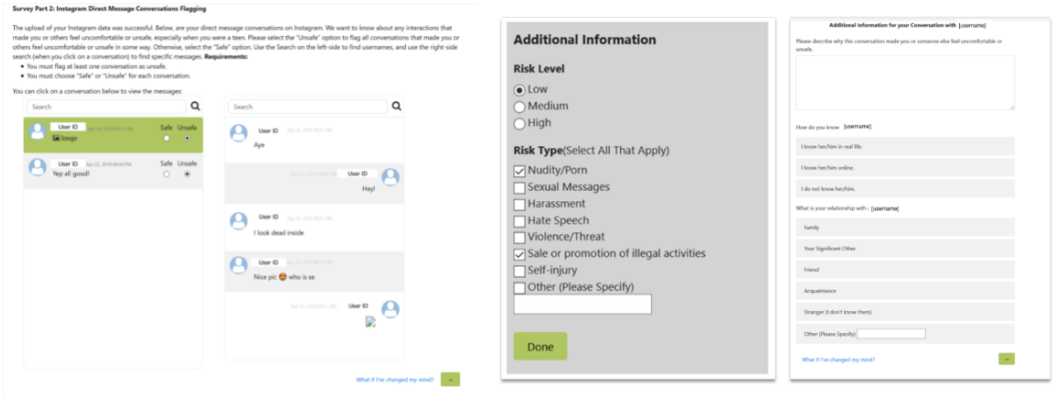

Fig. 1. Screenshot of (a) Youth DM Risk Annotation Interface (i.e., Safe/Unsafe at Conversation-Level) (b) Youth Risk Context Annotation Interface (e.g., Risk Level, Risk Type, Relationship Type).

annotation and data verification process were completed, participants were compensated with a $50 Amazon gift card for their data and time.

3.2 Risk-Flagging Annotation Process

We presented participants' Instagram DMs back to them in reverse chronological order, so they could review their conversations starting from most recent interactions, similar to how it is displayed on Instagram. We asked them to flag conversations that made them feel uncomfortable or unsafe as 'unsafe' and the rest of conversations as 'safe' as displayed shown in Figure 1a, left side. To make the process easier for participants that were 18 to 21 years old, we only displayed conversations during the time they were a teen (13-17 years old) to flag. For each conversation that was flagged as unsafe, we then asked participants to flag risk at the message-level, then identify the risk level and type (definitions available in Appendix A) of each message, as shown in Figure 1b, right side. We leveraged risk categories based on Instagram's reporting feature [3], so our categories matched with what participants commonly experienced on Instagram. However, we gave them the option to select "Other" to specify risk types that fell outside of the ones already specified. For this paper, we included risk flags specific to our pre-defined category of "Sexual Messages or solicitations" defined for participants as "Sending or receiving sexual messages. Being asked to send a sexual message, revealing/naked photo." in our analysis. Although we provided pre-defined risk types, we explained to the participants that unsafe interactions are not limited to these categories and they should self-assess the situations that felt unsafe to them. Please note that we used the term "risky" for uncomfortable or unsafe conversations throughout the paper. Participants were also asked to provide more contextual details about each unsafe conversation; for instance, whether the other party in the conversation was an acquaintance, a friend, a significant other, or a stranger. Because of our understanding that pre-existing relationships affect responses in online sexual experience incidents, we considered the knowledge of this relationship relevant to these risk situations [59].

3.2.1 Data Verification Process. We took special precautions to make sure that the data we collected was high quality. A team of researchers verified the data and compensated participants who passed a battery of quality checks. To this end, our data verification team removed participants who did not meet the eligibility criteria (e.g., did not have at least two unsafe conversations with back-and-forth messages), took unrealistically little time to complete the study, did not answer the survey attention

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

check questions, and whose Instagram data file seemed to be a fictitious account (e.g., lack of historical data, lack of face validity of conversations).

3.2.2 Data Privacy and Ethics. Our dataset contained highly sensitive and intimately personal information, which was why it was very important to preserve the confidentiality, privacy, and security of our participants. To do this, we first obtained Institutional Review Board (IRB) approval for our study. We disclosed our status of mandated child abuse reporters and our obligation to report child pornography (i.e., any nudity of a minor under the age of 18) to the proper authorities. Therefore, we gave strict warnings to refrain from uploading any digital imagery involving nudity of a minor and gave step-by-step instructions on how to remove such data prior to uploading data to our system. We also obtained a National Institute of Health Certificate of Confidentiality to further ensure participant privacy and prevent the subpoena of the data during legal discovery. While performing our data analysis, we also took several precautionary measures. We refrained from using any cloud-based services when analyzing our data and restricted data storage to university approved, secured devices. We also provided mental health support, such as adequate breaks for students who helped verify the data as some of the content could be triggering or explicit. For our participants, we also included "Help Resources" (e.g., sexual victim and suicide prevention hotlines) to be accessible during the study. In reporting our results, we removed all personally identifiable information and paraphrased quotations as recommended by Bruckman [16]. Since our data was private, they were not indexed by search engines to be reverse searchable; otherwise, we would have used more rigorous methods (c.f, [61, 62]) to further disguise the quotations presented in this paper from search engine discovery. To read more about the dataset creation and ethical challenges please refer to [58].

3.3 Participants Demographics

Our study was comprised of 150 participants between the ages of 13 to 21 (Av.=16 yrs, Std.=6.2). To recruit a diverse subset of participants, we promoted our study on social media and contacted more than 650 youth-serving organizations. Our participants were mostly female (Approx. 69%) with 21% identifying as males, 9% non-binary and the rest of the participants choosing not to provide their gender. The majority of our participants were heterosexual or straight (47%); however, a relatively large percentage of our participants identified as bisexual (29%), homosexual (11%), or preferred to self-identify (13%). The race distribution of our participants was as follows: Caucasian/White (41%), African-American/Black (20%), Asian or Pacific Islander (14%), and Hispanic/Latino (6%), and 19% belonging to mixed races or who preferred not to self-identify. We had representation from the following states: Florida (15.8%), California (12.5%), Indiana (2.6%), and 28 other U.S. states. Participants reported that they used Instagram several times a day (51%), every day or almost every day (22%), several times an hour (19%), once or twice a week (4%), less than once a month (2%), and less than once a week (1%). Table 1 listed the distribution of the safe (N=13,610) and unsafe (N=2,033) conversations by participants' gender, age, and relationship type.

3.4 Characteristics of the Instagram Data

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

Table 1. Proportion of safe ($N = 13,610$) and unsafe ($N = 20,33$) conversations across contextual factors

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

as high risk. Most participants said that the sexual unsafe conversation was with someone they did not recognize 73%, some knew them from real life 14.4%, and some knew them online 10.1%. Table 1 showed that most of the unsafe sexual online conversations of youth were with strangers. Out of these 326 conversations, 26 conversations were group conversations.

4 METHODS

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

4.1 Data Pre-processing and Preparation

During the data pre-processing phase, punctuation marks, hyperlinks, stop words, non-latin words, single/numeric characters, and conversations that had less than three words were removed. Emojis were converted to their associated word representations through the demoji Python library[4] to conserve the semantic meaning depicted through emojis.

To train our conversation-level classifiers, we first created a dataset that had a 50-50 split between the conversations that were labeled with sexual risk and those that were not labeled with any online risk. In applications such as risk detection where datasets are usually imbalanced toward having more safe samples rather than unsafe samples, it is common to balance the data between classes [18, 74]. We used random under-sampling to reduce the number of safe conversations to gain an equal number of class samples and create a balanced dataset. After the data cleaning, our dataset contained 264 sexual unsafe and 249 safe conversations (Total=513).

For classifying the risk level of unsafe sexual messages we had 3 classes with uneven samples (Low= 136, Medium= 65, High= 33). We encountered two issues with this data, first uneven samples and second lack of data. Thus, we oversampled the unsafe sexual messages with risk levels to the class with the highest instances (N=136) to create a balanced dataset. We used the RandomOverSampler[5] library which over-samples the minority classes by picking samples at random with replacement. After oversampling, we had a balanced number of samples for each class (Low=136, Medium=136, High=136), but still did not have enough data for classifiers to perform

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

[5]https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.RandomOverSampler.html#imblearn.over_sampling.RandomOverSampler

well. Therefore, we used a Contextualized Word Embedding augmentation by NLPaug Python library [6] Contextual Augmentation called BertAug [43] and doubled the number of instances in each class (N=272). Classic word embeddings might not fit some scenarios since they use a static vector to represent the same word with different meanings. Meanwhile, contextualized word embeddings consider surrounding words to generate a vector under a different context to solve this issue [7]. BertAug provided insertion which is predicted by the BERT language model which is better than picking one word randomly [43]. Also substitution uses surrounding words as a feature to predict the target word. We leveraged augmentation because it is useful in several aspects, including minimizing label effort, lowering the usage of real-world data in sensitive domains, balancing unbalanced datasets, and increasing robustness against adversarial attacks [12].

Finally, we used the PAN12 dataset [37], which was created for the sexual predation detection competition, to compare the performance of our classifiers to the state-of-the art in sexual risk detection as a baseline. This dataset included 66,927 non-predatory conversations (from 97,695 users) and 2,016 predatory conversations (142 users) for training, and 15,5128 conversations (218,716 users) and 3,737 predatory conversations (254 users) for testing. The predatory conversations were gathered from the PJ dataset including chats of volunteers acting as teens with convicted sex offenders while the non-predatory/non-sexual samples were provided from publicly available Internet Relay Chat logs, which mainly contain chats about computer and web technologies [37].

4.1.1 Feature Engineering. We developed five categories of features for our traditional supervised learning models. We used the features below to build sexual risk detection classifiers for detecting unsafe sexual conversations (RQ1). For detecting unsafe sexual messages and their severity (RQ2), we trained the traditional models with combination of the five feature types. For feature analysis, we trained and tested the sexual risk conversation classifiers with each category separately, as well as using all five together (RQ3). We used the flagged messages of the participants as ground truth to train these models. Each conversation/message was represented as a vector where each element was the value of one feature. In the following list, we describe the features that compose each category in more detail:

- **Contextual Features (Age, Gender, and Relationship):** We acquired age and gender of the participants from our survey questions. We examined them as features (3 options for gender and an integer for age) for our model since many empirical studies emphasized the role of age and gender in online sexual risks [21]. For the unsafe conversations, participants were also asked the nature of the relationship between themselves and others involved. Based on this participant annotated data, we trained a Convolutional Neural Network (CNN) model (Avg. AUC=0.90) based on concurrent work [anonymized for review] to machine label the safe conversations for relationship type (i.e., stranger, acquaintance, friend, family, significant other). The relationship feature was a categorical number representing the relationship type for each conversation.
- **Psycholinguistic Attributes (LIWC):** LIWC is commonly used to obtain psycholinguistic features embedded in text as well as to quantify meaning across various dimensions [52]. Based on prior work [20], we selected 50 categories spanning across *affect, cognition and perception, interpersonal focus, temporal references, lexical density and awareness, biological concerns*, and *social/personal concerns* and used them as features. To calculate each feature, we normalized the word counts related to each category by the length of the conversation.
- **Sentiment:** Emotion was represented as a sentiment score, which were extracted through Stanford CoreNLP's deep learning tool [49]. The tool gave us a single label that indicated

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

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

- **Term Frequency-Inverse Document Frequency (TF-IDF):** TF-IDF scales down the term weights of terms with high collection frequency by reducing the weight of a term by a factor that grows with its collection frequency [66]. We defined each conversation as a document, and calculated the 25 words with the highest TF-IDF in all conversations. We then used these 25 words as features, calculating the normalized count for them in each conversation.
- **Sexual Lexicon:** To capture domain-specific signals as features, we used a lexicon developed in prior work [59] including 98 words. For each of the words, we use its normalized count in a conversation as a feature.

4.1.2 Machine Learning Models. We chose Linear Support Vector Machine (SVM), Random Forest (RF), and Logistic Regression (LR) as traditional classification approaches. Next, we implemented an end-to-end Convolutional Neural Network (CNN) – a technique which has shown promising results for text classification [42] in addition to image recognition tasks [19] in recent years. This allowed us to compare the results of the CNN model with the traditional models. In order to convert conversations to vectors of tokens as input of the CNN model, we used the Keras Tokenizer Python library[8]. We also experimented using pre-trained GloVe [53] to convert text to word embeddings, which capture the semantics and syntax of words in text. We then built a CNN model that aims to predict whether a conversation/message is sexually unsafe and risk severity of a message. We used participant flagging or annotation as ground truth to train and evaluate this classifier.

4.1.3 Evaluation. We used the average accuracy of the models, standard deviation of the accuracy, F1-measure, area under the receiver operating characteristic curve (AUC), and class-specific precision and recall to evaluate our models on the test sets. We used grid search and stratified k-fold cross-validation ($k = 10$) to tune the hyper-parameters during the training and validation phases. While the accuracy and F1 scores return the general performance of the models, precision and recall of each class and AUC provide more detailed insights.

4.1.4 Unpacking the Quantitative Results Qualitatively. To understand the nuance as to why some conversations were classified as safe or unsafe by our quantitative models, we examined strongest contributing features for traditional models. To do this, we first performed Chi-square (χ^2) tests [68] to identify the statistically significant differences between top features in the Unsafe Sexual/Safe conversations and used that to further unpack the nuance found within our qualitative results. We examined top SVM features (which were the LIWC categories and the contextual features), as the SVM coefficients were interpretable. For features where statistical significance was reached, we dove deeper with qualitative content analyses [36] to better understand why and how these features contributed to the ML performance. Lastly, we qualitatively analyzed misclassified samples to further dive into why errors occurred.

5 RESULTS

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

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

Table 2. Model performance across different feature sets for traditional models and CNN performance.

Features	Classes	Prec.	Rec.	F1	AUC	Accr.
Linear SVM						
Combined + Contextual	Sexual	0.77	1.00	0.87	**0.84**	**0.85**
	Non-sexual	1.00	0.68	0.81		
Combined	Sexual	0.73	0.65	0.69	0.70	0.71
	Non-sexual	0.69	0.77	0.73		
LIWC	Sexual	0.83	0.77	0.80	0.76	0.77
	Non-sexual	0.70	0.76	0.73		
Sentiment	Sexual	0.62	0.53	0.57	0.57	0.57
	Non-sexual	0.53	0.62	0.57		
TF-IDF	Sexual	0.73	0.77	0.75	0.67	0.69
	Non-sexual	0.63	0.57	0.60		
Sexual Lexicons	Sexual	1.00	0.05	0.10	0.53	0.49
	Non-sexual	0.47	1.00	0.64		
Random Forest						
Combined + Contextual	Sexual	0.86	0.93	0.89	**0.88**	**0.88**
	Non-sexual	0.91	0.84	0.87		
Combined	Sexual	0.90	0.70	0.79	0.81	0.81
	Non-sexual	0.74	0.92	0.82		
LIWC	Sexual	0.74	0.74	0.74	0.68	0.69
	Non-sexual	0.62	0.62	0.62		
Sentiment	Sexual	0.83	0.61	0.70	0.71	0.69
	Non-sexual	0.59	0.81	0.68		
TF-IDF	Sexual	0.83	0.81	0.82	0.78	0.79
	Non-sexual	0.73	0.76	0.74		
Sexual Lexicons	Sexual	0.75	0.41	0.53	0.63	0.64
	Non-sexual	0.60	0.87	0.71		
Logistic Regression						
Combined + Contextual	Sexual	0.76	0.96	0.85	**0.82**	**0.83**
	Non-sexual	0.94	0.68	0.79		
Combined	Sexual	0.78	0.67	0.72	0.73	0.73
	Non-sexual	0.69	0.80	0.74		
LIWC	Sexual	0.77	0.77	0.77	0.72	0.73
	Non-sexual	0.67	0.67	0.67		
Sentiment	Sexual	0.59	0.53	0.56	0.54	0.54
	Non-sexual	0.50	0.56	0.53		
TF-IDF	Sexual	0.79	0.61	0.69	0.68	0.67
	Non-sexual	0.57	0.76	0.65		
Sexual Lexicons	Sexual	0.89	0.42	0.57	0.68	0.66
	Non-sexual	0.58	0.94	0.71		
CNN						
Language Tokens	Sexual	0.80	0.92	0.86	**0.88**	**0.89**
	Non-sexual	0.95	0.86	0.90		

5.1 Conversations-level Sexual Risk Detection (RQ1)

We implemented and evaluated multiple classifiers detecting sexual risks at the conversation-level, using the 476 conversations as training and test datasets. Minimal pre-processing of data improved performance for the traditional models. For instance, while we experimented with lemmatization and spellchecking, traditional models using these resulted in relatively poorer performance because pre-processing removed contextual and linguistic style information from the original conversations. Further, for the CNN model, recall that we experimented using pre-trained GloVe [53] to convert text to word embeddings. However, the CNN performed better with the Tokenizer, because GloVe had a disadvantage with out-of-vocabulary words from the corpus that it was pre-trained with.

Table 2 summarizes the performance metrics of the traditional machine learning models with different feature sets, and the performance of the CNN classifier on the IGDD private dataset. Overall, we found that the RF model with combined plus contextual features outperformed other traditional classifiers with an AUC=0.88 and accuracy=0.88, which was similar to the CNN model (ROCs displayed in Figure 2). Also, it achieved high class specific precision and recall. For the sexual class, the recall (0.93) was higher than precision (0.86) and for the non-sexual class the precision (0.91) was higher than the recall (0.84).

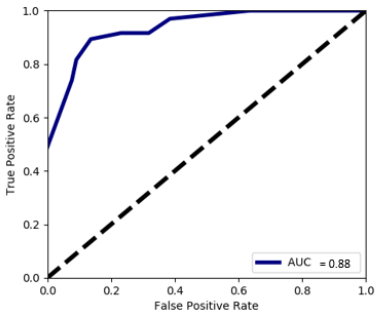

Fig. 2. CNN Sexual Risks Conversation Classifier ROC

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

5.1.1 Comparison with PAN12 data as a Baseline. To compare the performance of the models with baselines, we trained and tuned the models on the PAN12 dataset[9]. Table 3 presents the results of the baseline models on PAN12 dataset. The data pre-processing and preparation was completed in a similar manner to the IGDD data as mentioned in section 4.1. We only included the results for the combined features in Table 3, since these features had the best performance for the PAN12 dataset compared to individual features similar to IGDD data. Contextual features were not available for PAN12 dataset to be used. The best performing model was RF (AUC=0.92, accuracy=0.92) which achieved higher performance than the first place models for the PAN12 competition [37], shown in the last row of the Table 3. Unfortunately class specific prevision and recall, AUC, and accuracy were not reported at PAN12 competition. Next, we investigated how the classifiers that were trained on public dataset (PAN12) performed on real-word private datasets (IGDD). We found that when

[9]https://pan.webis.de/clef12/pan12-web/sexual-predator-identification.html

Table 3. Models' performance on PAN12 dataset as a baseline for public datasets.

Classifiers	Classes	Prec.	Rec.	F1	AUC	Accr.
Linear SVM	Sexual	0.86	0.83	0.84	0.85	0.85
	Non-sexual	0.84	0.86	0.85		
Random Forest	Sexual	0.92	0.92	0.92	**0.92**	**0.92**
	Non-sexual	0.92	0.92	0.92		
Logistic Regression	Sexual	0.86	0.82	0.84	0.84	0.84
	Non-sexual	0.83	0.86	0.85		
CNN	Sexual	0.89	0.85	0.84	0.91	0.86
	Non-sexual	0.86	0.86	0.86		
Best result of PAN12 Competition		0.98	0.78	0.87	-	-

Table 4. Sexual Risks Message Classifiers' Performances.

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

these classifiers were tested on IGDD, the performance was reduced to AUC=0.46 accuracy=0.46 for RF (highest performance). This means that classifiers performed poorly and are not suitable for predicting the sexual risks within private conversations.

5.2 Message-Level Sexual Risk Detection (RQ2)

In this section, we presented the results of the classifiers for detecting sexual risks in messages and then detecting risk severity of a given message. Given that the traditional models using combined features plus contextual features performed the best for our conversational-level classifiers (RQ1), we also used this approach when building out message-level classifiers (RQ2). Moreover, we tested the message-level classifiers with each feature and combination of features, and combined features plus contextual features performed the best, so it was only included for the purpose of parsimony. This enabled us to directly compare the performance of our message-level classifiers to our best performing conversation-level classifiers.

5.2.1 Binary Classification. While our conversation-level classifier performed well, message-level detection was necessary for real-time risk detection and mitigation. Therefore, we trained classifiers for detecting sexual risks at message-level to compare the results with the conversation-level classifiers (displayed in Table 4). CNN model outperformed the traditional models with AUC=0.85 and accuracy=0.82. CNN and RF performed better in conversation-level rather than message-level. Though, SVM and LR were slightly better in message-level compared to conversation-level.

5.2.2 Classification by Risk Level. Since participants flagged the risk severity levels (i.e., low, medium, and high) of sexual messages, we trained classifiers to detect the risk severity level of

Table 5. Message Risk Severity Classifiers' Performances

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

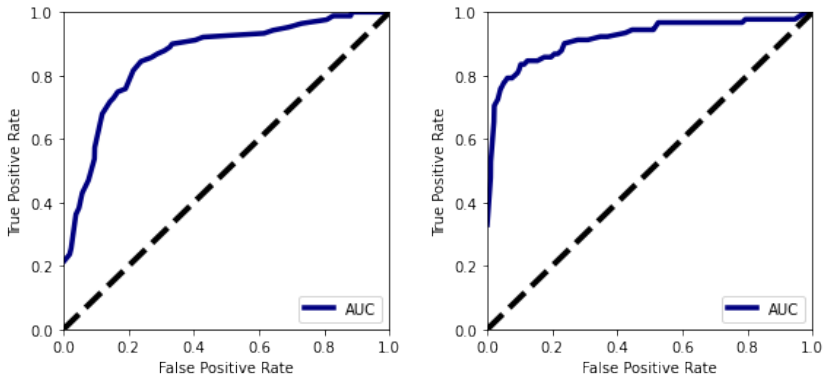

Fig. 3. (a) Sexual Risks Message-level CNN ROC (b) Risk Severity Level CNN ROC.

messages in unsafe conversations. Identifying risk levels can be helpful in the process of real-time risk mitigation. For classifying risk levels, we filtered the original dataset to include the unsafe sexual conversations and trained the risk level classifiers for messages within these conversations flagged by participants for low (N= 136) and combined medium and high (N=98) risk levels (due to the smaller numbers). We randomly selected an equal number of messages from safe conversations (N=234) to classify the messages into safe, low, and medium-high risk levels. We used oversampling (explained in the Method section) to make the classes with lower number of samples equal to the larger sample (each class N=234). We used this approach to make a balanced dataset since participants flagged fewer number of messages as high and medium risks compared to low risk. We trained traditional classifiers with combined features and the CNN with language tokens classifier and compared the results demonstrated in Table 5. The results of the unsafe sexual message-level classifiers are shown in Table 5. The CNN classifier outperformed SVM, RF, and LR and resulted in AUC=0.88 and accr=0.82 evaluated by 10-fold cross validation.

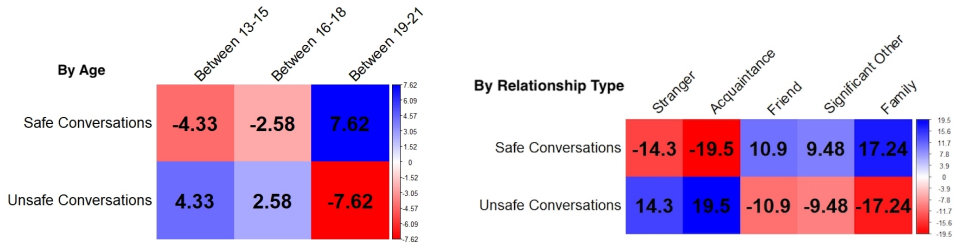

Fig. 4. Correlation matrix of Pearson's standardized residuals between (a) age groups (b) relationship types and their safe/unsafe conversations.

5.3 Contextual Features and LIWC Analyses (RQ3 a)

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

5.3.1 Contextual Features (Age, Gender, and Relationship Type). Since the combination of features plus contextual features yielded the best performance for the models comparatively, we further analyzed this data to uncover patterns. First, we dug deeper into the contextual features.

For age, a χ^2 test indicated a significant difference between age groups ("Between 13-15", "Between 16-18", "Between 19-21") and their conversation flagging (safe / unsafe) behavior $\chi^2(df = 2, N = 15,547) = 63.33, p < 0.001$. Post hoc testing revealed that younger teens (ages 13-15) were significantly different from older adolescents (ages 16-18) ($p = 0.02$) and young adults (ages 19-21) ($p < 0.001$). There was also a significant difference between older teens (ages 16-18) and young adults (ages 19-21) ($p < 0.001$). The proportions of safe and unsafe conversations as shown in Figure 4 indicated that young adults (ages 19-21) were more likely to flag their conversations as safe, while younger teens (ages 13-15) and adolescents (ages 16-18) were more likely to flag their conversations as unsafe.

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

For relationship type, a χ^2 test showed a significant difference between the relationships types (Stranger, Acquaintance, Friend, Significant Other, Family) and the conversations flagged as safe versus unsafe ($\chi^2(df = 4, N = 15,547) = 882.37, p < 0.001$). According to the post hoc analysis, there was a significant difference between Friend and Significant Other ($p < 0.001$). The post hoc test also found significant differences between Strangers and all other relationship types (Acquaintances ($p < 0.001$), Friends ($p < 0.001$), Significant Others ($p < 0.001$), and Family ($p < 0.001$). For Acquaintances, we found significant differences with Friends ($p < 0.001$), Significant Others ($p < 0.001$), and Family ($p < 0.001$). Regarding Family, there were significant differences from Friends ($p < 0.001$) and Significant Others ($p = 0.001$). Overall, the proportions of the safe versus unsafe conversations showed that participants were most likely to report having safe conversations with family members, friends, and significant others; meanwhile, unsafe conversations were most likely with strangers or acquaintances, as shown in Figure 4.

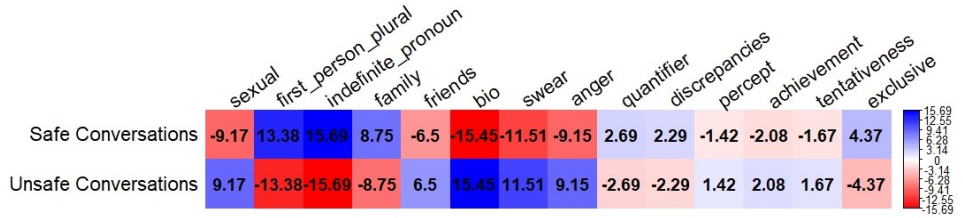

Fig. 5. Correlation matrix of Pearson's standardized residuals between top LIWC predictive features for SVM and Unsafe Sexual Conversations (Red) and Safe Non-sexual Conversations (Blue).

5.3.2 LIWC Categories. A benefit of linear SVM was that it is an interpretable model to find the most contributing features by looking at the model's coefficients. Therefore, we chose the next best performing feature based on the SVM's AUC to further examine the linguistic contributing factors; as the Linear SVM trained on the LIWC had the highest AUC, we looked at the top 15 LIWC psycholinguistic categories in terms of their importance given by the model, as shown in Figure 6. To further confirm the significance of the LIWC categories, we performed a χ^2 test which demonstrated a significant difference between the LIWC categories (top LIWC predictive features for SVM) and the conversations flagged as safe versus unsafe (χ^2 $(df = 13, N = 147,797) = 1016.7, p < 0.001$) with the standardized residuals illustrated in Figure 5. For instance, first person plurals category, such as *"we are,"* which could signal a sense of group identity or togetherness [73], found to be one of the top predictive features for safe conversations. For example, in a conversation that a 16-year-old female participant had with her friend, they were talking about joining a club, and they used first person plurals frequently to reference their collective action:

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

In addition, the indefinite pronouns, such as *"It," "it's," and "those"* appeared also as a top predictive category. These linguistic cues tended to show more interest in objects and things [73]. For example, in a conversation that a 15 year-old female participant described as an *"unwelcome advance:"*

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

As illustrated above, it was less likely for participants to use collective, first person plural language when they sought to separate themselves from offensive behavior, and offenders used indefinite pronouns to objectify their victims. For LIWC social processes category, safe conversations most likely included words about family, yet unsafe sexual conversations included linguistic cues about friends. Safe conversations were mostly daily interactions with known others, where it was commonplace to make a reference to a family member. For instance, in a safe conversation between a 16-year-old female and her friend, they talked about regular things that happen with their family:

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

In unsafe conversations, however, there were more friend-related words, even when the conversation was with a stranger. This was often because the stranger was trying to become familiar with the participants or propositioned them to a *"friend with benefits"* or *"sugar baby"*. In the following unsafe conversation, a stranger sent an unwanted sexual solicitation of this nature. Our 18-year-old

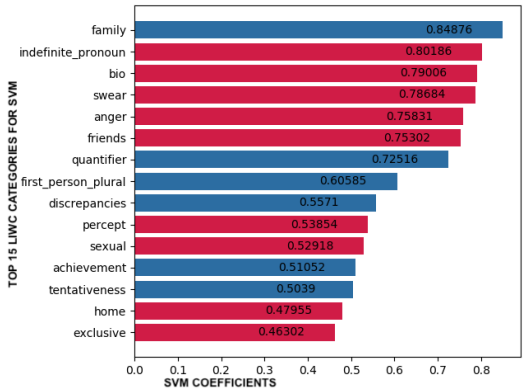

Fig. 6. Top LIWC predictive features for SVM with their absolute coefficients for Sexual (Red) and Non-sexual classes (Blue).

female participant did not respond to the unwanted advance. One could see that the text also included indefinite pronouns:

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

For affective processes, which was a LIWC category for emotionality [73] *"Happy," "cried," and "abandon"*, unsafe sexual conversations most likely contained negative emotions, such as anger and swear words. For instance, there were lots of profanity, sexual words, and negative emotions in group chats, usually among males. An 18-year-old male participant described one conversation that made him feel uncomfortable or unsafe when he was 16-year-old as:

> **Participant Description:** *They were sexual messages and messages about self harm sent to me at a young age by people i did not know in real life.*
>
> **Other Person:** *Guys I'm so stressed. History is fucking me in the ass!*
> **Other Person 1:** *Thought that was my job*
> **Other Person 2:** *I wanna kill myself.*
> **Other Person 3:** *I'm sick. Wtf was I doing before I went to bed. Idk probably watching porn or some shit.*

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

> **Other Person:** *Hi beautiful My ex baby sent me feet pics, lingerie pics. Texted regularly and FaceTimed in exchange for a weekly $200 allowance and a $800 monthly shopping voucher It's more like a companionship type relationship while I'm away. nothing sexual Would you be open to such?*

Our participant contextualized the exchange as *"I do not know who this person is. It seems likely to be a bot or phishing scam."*

In addition, sexual conversations more likely included LIWC perceptual processes such as *"seeing"* which referred to emotional and physical sensations. In contrast, safe conversations mostly included

words from cognitive processes, such as discrepancies (*e.g. "should," "would," or "could"*) and tentative (*e.g. "Maybe," "perhaps," or "guess"*).

5.4 Error Analysis (RQ3 b)

Next, we looked into specific prediction instances to provide more insights on the factors that contributed to misclassifications. We qualitatively investigated the linguistic style used in the misclassified conversations for the SVM and CNN models, and then misclassifed instances for the SVM messages' risk level classifier.

5.4.1 False Negative (FN) Conversations. We performed a recall-centric analysis of the FNs for the conversation level classifiers. First, all of the common FN samples between CNN and the SVM models included media, which our text-based models were not able to identify. An example of such instances was a conversation in which someone shared a sexually explicit drawing with a 22-year-old female participant, who described it as *"drawings based on real nude photos of others that I did not consent to seeing"*. The participant in the conversation asked *"all based on real nudes I'm guessing?"* Although this sentence included *"nudes,"* it was paired with the word *"guessing"* from the *tentativeness* category; therefore, the conversation was classified as safe.

The SVM's FN samples were often instances where the participant was added to a group chat with a sexual title and sexual links shared between group members. Since these conversations were short and only included sexual links/media and a sentence naming the group to something sexual, such as *"contact named the group My best nude pic's."*, the model was not able to identify it correctly. In these instances, participants often left the group immediately, which was why the conversations were very short. For instance, a 15-year-old female was added to a group chat with porn links, and she described that *"i kept leaving the chat and people kept adding me back in"*. We inspected similar group chats that contained longer messages, where the model was able to identify them correctly. Some conversations did not have enough linguistic cues and were based on some prior context not present in the conversation. For example, a 19-year-old female received a compliment on an Instagram story saying "Cute I mean, you always are, but here particularly so" and she responded "Haha [smiling face, smiling face with heart] thank you!" Since the conversation included positive emotions and appreciation, which were related to safe conversations, the SVM classified it as non-sexual conversation. But the participant flagged this conversation with her classmate as sexual and provide more context as, "I met him once in school, and during this time, he had begun to stalk me in person. He started to comment very sexual things about me, and it made me feel very uncomfortable." Overall, the SVM model had more FNs compared to the CNN model and were not able to identify short conversations with minor sexual content such as short solicitations including greetings and compliments or flirtations from strangers.

5.4.2 False Positive (FP) Conversations. Now we discuss a precision-centric analysis based on FPs. Common FPs between the CNN and SVM models were short conversations included second person pronounces and included either a request such as *"hi can you tell me how to do a uu move"*, compliments such as *"omg i can not even begin to describe how amazing of a writer you are... i can not believe a person could do this through a book i am amazed by you"* or group conversations around topics of video games to sports, but including profanity and sexual words.

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

conversation included LIWC categories, such as swear, money, and sexual words that were more often associated with unsafe sexual conversations:

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

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

Similar to SVM, most of the FP instances of the CNN were short conversations with only 1-3 messages that included an Instagram link or images. The common characteristics of these messages were that they included one of the following; a request statement such as "ik this is not the best place to meet people but i am actually really nice and easy to talk to ... but i understand if you do not feel comfortable", a compliment "ur so pretty lol", appreciation "thank you", or love emojis. As making requests and complimenting are from the tactics and the stages for sexual solicitations/grooming [51], the CNN model classified those as unsafe conversations. Naturally, the adolescent respond to those requests by saying "thank you" and that's mostly why the appreciation words were also misclassified.

5.4.3 Messages' Risk Level Misclassification Analysis. In this subsection we present the analysis of misclassified instances for Messages' risk levels for the SVM classifier. Most of the misclassified *mid-high risk level* messages were classified as low risk, which they were mostly group name changes such as "a contact named the group my nude no one under eighteen pics", or included compliments such as "hey i am glad you followed me i really thought you were beautiful lol" and sexual words such as "give me naked" (N=2). The mid-high risk messages misclassified as safe mostly were short messages that did not included the context of the conversation (belonged to a long conversations with media shared) such as "you want' or 'i need it' or they included compliments. Also it was hard for the classifier to catch the idioms or expressions for words expressed in sexual ways without their literal meaning such as "can you want see my lancer 22 cm". Overall, *Low risk level* had the most misclassified instances, which were mostly classified as mid-high risks. These instances included sexual messages but were mostly sarcastic conversations for fun, so their literal meaning might be more of a high risk than low risks such as "you got a girl pregnant." There were a few *safe messages* misclassified, in which all of those were very short messages such as 'is it like tapenade' or "sorry i haven't talked". These were misclassified because of being very short and having some unsafe cues such as indefinite pronouns.

6 DISCUSSION

In this section, we discuss the key implications from our findings based on our three overarching research questions. In summary, our work takes a HCML approach to advancing computational approaches on sexual risk detection for youth. First, we constructed an ecologically valid dataset that is composed of private conversations donated by youth participants. With self-reported labels from participants, we not only detect sexual predators but also assessed the survivors' perspectives of the sexual risk experience. This is a significantly different goal than attempting to identify sexual predators. Built upon this ecologically valid dataset and labels, this paper also incorporates human-centered features in developing an automated sexual risk detection system. Next, we qualitatively analyzed instances of our top performing features to shed light on the sexual risk experiences of the youth participants in our dataset. Specifically, we conducted a feature analysis to quantitatively and qualitatively understand how our feature set not only contributed to our prediction accuracy,

but also to better understand the experiences of our participants. Finally, we conducted an error analysis to pinpoint areas of weakness that should be addressed in the future.

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

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

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

In the case that an algorithm such as ours is embedded in a criminal justice system to identify sexual predators, the legal system typically puts the burden of proof on the prosecution [10] as those who are labeled as sexual predators face criminal charges. Therefore, a model with high precision would align with the goals of the system as it aims to avoid wrongful accusations [14]. However, we advocate that our models instead be used for the purpose of risk prevention (rather than prosecution after-the-fact) and embedded directly within the social media platforms that put youth at-risk, especially with the heightened concern about the well being of youth on Instagram in recent news [11]. This recommendation is consistent with recent U.S. legislation [12] that passed to fight online sex trafficking and stop enabling sex traffickers (FOSTA-SESTA Acts) by making online platforms accountable for user-generated content that promotes sexual violence. To maximize risk prevention, therefore, models with a high recall that aim to prioritize providing support to any potential victims at the cost of false alarms would be suitable for this purpose. Yet, the goals of the primary stakeholders (i.e., adolescents and young adults), in this case, are less clear. Would young social media users get annoyed by too many false warnings of possible unsafe sexual content (similar to what has been learned about overzealous security warnings [44])? Or possibly, would youth benefit from being made aware of the implicit sexual undertones in their conversations (even within their own messages) and benefit from receiving just-in-time support for how to handle such risky situations? Such support may be particularly effective for young adolescents (ages 12-15), as stigma acts as a potential barrier when they seek help [69]. Furthermore, raising risk awareness could help teens identify and appropriately respond to these risky experiences before they escalate to emotional or physical harm [38, 79]. More research is warranted to better understand stakeholder

[10] https://www.law.cornell.edu/wex/burden_of_proof
[11] https://www.cbs8.com/article/news/health/whistleblower-brings-attention-to-facebook-and-instagram-affecting-young-peoples-mental-health-san-diego-doctor-psychiatry/509-e5d5c810-8491-4186-b6b4-6328ff82fa1f
[12] https://www.vox.com/culture/2018/4/13/17172762/fosta-sesta-backpage-230-internet-freedom

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

6.1.2 Conversation-level vs. Message-level Trade-offs. Another computational trade-off is at the unit and/or level of analysis in which classification occurs. While our conversation-level classifier performed the best, message-level classification is necessary for real-time risk detection and mitigation. While risk detection at the conversation-level provides more contextual and linguistic cues for more robust classification, detecting sexual risk or after-the-fact based on entire conservation may come at the risk of being too late. Post-hoc conversational-level detection may be beneficial when attempting to apprehend sexual predators [33], but it does little to protect youth from becoming victimized in the first place [60]. Prior research on the Perverted Justice dataset were successful in terms of predicting predatory conversations after-the-fact that risk happens [9, 13, 33], but were not able to detect predatory lines that make a conversation risky [37]. Therefore, we propose a few ways to balance these trade-offs. One option would be to implement sexual risk classification at both levels (conversation and message) for a two-level classification system. The multi-tiered system would more readily be able to support real-time risk detection for mitigation purposes, while the conversation-level detection system could ensure the overall robustness by keeping context intact. Another alternative would be leveraging models that take time dimension into account for predictions that they make, to facilitate more robust risk detection in real time. Having different sexual risk levels introduced a tangible way to allocate resources for treatment, support, and prevention of online risks in an effective manner. The existence of risk in social media could be initially seen as a binary classification task; however, this provides a foundation for future work to study the difference in linguistic styles of risk severity to further aid with providing support to the victims before the high risk levels happen.

6.2 Understanding the Private Digital Lives of Youth (RQ3)

Through our human-centered lens, we uncovered important insights about the contextual and linguistic features that were indicative of the sexual risk experiences of youth. Importantly, we found that the age, gender, and the relationship between our participants and others, plus combined features, resulted in the best feature performance compared to more traditional NLP approaches (e.g., Sentiment, TF-IDF). This highlights the importance of including contextual information about people and the relationships between them, rather than relying on linguistic cues alone. We based these contextual features on empirical evidence from prior works [5, 11, 59] that found that gender, age, and relationship context were important factors associated with youth exposure to sexual risks. Past studies have uncovered the importance of relationship context in identifying the riskiness of youth online sexual interactions [25, 63, 75]; yet until now, the computational sexual risk detection has yet to leverage this knowledge in a meaningful way [28]. Additionally, while these prior works were primarily based on the self-reports of youth, our work is the first to triangulate these claims based on social media trace data from youth.

6.2.1 Importance of Interpretability. Our results also highlight the importance of how designing interpretable models using evidence-based research can be superior to black-box approaches. Deep learning models, which mainly prioritize accuracy at the cost of transparency and explainability [55], make it hard for the human users to make sense of *what* and *why* certain features are important. In contrast, our traditional ML models outperformed the deep learning models and added significant value in that they helped us gain deeper insights in the sexual risk experiences of our youth participants. Models' interpretability is defined in the literature of the explainable AI as "the ability to explain or to provide the meaning in understandable terms to a human" [8]. Interpretability of a model comes from the design of the model and it includes the human comprehension of models'

parameters. For example, a RF model would be represented as trees, SVM as the hyperplane plot with dots annotated with phrases, and CNN as sparse connected convolution matrix created in each layer[27]. The importance of interpretability of ML models have long been advocated by scholars in the HCML area; researchers have argued that interpretability allows us to understand the impacts of machine learning models on the stakeholders, to think about existing challenges and solutions to making the models' results more human-centered, and to establish the interpretations of what each model does [10, 41, 56, 60]. Comprehending the transparency of the models to see how the model is based on human-centered approaches [50] is crucial when reviewing the computational risk detection models. Examining each component of the models such as types of algorithms, feature sets, and parameters all contribute to understanding the meaning of the models [35].

6.2.2 The Language of Sexual Victimization. From a more human-centered perspective, our results shed light on the language used to sexually groom, objectify, and victimize youth in private online spaces, which has important implications for both social and computational science, as well as for victim advocacy. In our RQ3 results, we observed that first personal plurals (e.g., "we", "us") were used in safe conversations to show collective purpose and togetherness, while indefinite pronouns (e.g., "it," "that") were more often used by perpetrators of sexual risk to objectify victims and by victims to put distance between themselves and others who made them feel uncomfortable or unsafe online. Further, sexually risky conversations were more likely to contain negative emotions, profanity, words descriptive of biological processes and body parts (e.g., "blood," "pain"), and sex (e.g., "horny," "love," "incest"). A key implication from these findings is that educational and awareness programs for sexual violence and sex trafficking prevention could leverage these insights in training materials that empower women and other vulnerable people (e.g., LGBTQ+) on the linguistic cues indicative of sexually risky dialogue, as well as effective strategies for taking protective measures against these advances when unwanted. A question raised through these insights, however, is whether sexual language used in the formation of *wanted* online romantic relationships [45] mirrors predatory language or is distinguishable from it. Therefore, we recommend that future research also study the language used when youth are forming healthy romantic relationships online to attempt to answer this unanswered question.

6.3 Implications for Design of AI Sexual Risk Detection Systems

Our deeper analysis of contextual feature differences sheds insight into how the thresholds for sexual risk detection algorithms might be optimized for different users. For instance, younger teens were more likely to flag unsafe conversations than young adults; therefore, the classifiers may be fine-tuned to be more sensitive (more tolerant to FPs) to sexual risks for younger users, while less sensitive (erring toward FNs) for young adults. Yet, while social media platforms typically request a users' age upon account creation, the more recent research on age verification [70] should be considered given youth are known to sometimes lie about their age when joining these platforms [30]. Similarly, algorithmic sexual risk detection systems could identify the nature of the relationship to make sexual risk detection more attuned to conversations between strangers and acquaintances, as opposed to friends, significant others, and family members. This would be an alternative to the stricter decision of Instagram and other social media platforms to ban strangers from having direct message conversations with minors who do not follow them [13]. Of course, such deploying AI in this way would require extensive user evaluation and design work to make sure that misclassifications based on age and relationship type did not unintentionally burden or harm end users in unexpected ways.

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

Integrating an ML system to automatically flag private conversations as sexually risk or not should be done with great care for social media users. In order to preserve users' privacy, light resource consuming local pre-trained models [46] could be implemented to detect online risks for youth, not sending all the private information to the cloud to be continuously monitored. We have seen how difficult the problem is when big platforms struggle to solve similar problems of harmful behaviour (e.g. hate speech, disinformation) on public discourse [14]. In a private setting, the problem is even more complex since there may be many subtleties in the communication between two individuals that would render the problem ill-defined for an ML system. Therefore, it becomes more important on what decisions will be made when such an instance is detected by the system. For instance, when a sexually risky instance is detected by the ML model, it should provide a suggestion to the youth user and the user should be able to provide feedback to the model and make the final decision. Using such human-in-the-loop approaches [4, 81] would help the user be in the control of the uncertainties of the real world. In addition, identification of sexually-risky content could get conflated with content that mentions sex in general such that in cases people discussing uncertainties around sexual identity, e.g., might be disproportionately harmed. For having scalable solutions in real world, handling possibility of false-positive/negative instances becomes of an important design decision. Since unnecessary suspicion on innocent conversations or missing instances of harassment can be unacceptable in the real world.

6.4 Limitations and Future Work

A key strength but also a limitation to the generalizability of our work is that we collected a difficult to obtain dataset of private Instagram conversations from youth (ages 13-21). Therefore, our results should only be generalized to this population. Further, since our analysis was based on Instagram, our results may be constrained by the unique affordances [77] of Instagram and not generalizable to other social media platforms. Therefore, more research is warranted on examining private conversations that occur on other platforms with other user groups to validate whether our sexual risk classifiers are transferable to those new contexts. We based our ground truth on the lived experience of the youth who participated in our study. However, for future work, we have two third-party annotators reviewing each conversation to flag them for additional risks that may not have been flagged by our participants. As future work, we will have clinical experts weigh in on the risk-flagged data from participants. We plan to do a comprehensive analysis of the discrepancies between annotations by participants and our third-party annotators. It would be interesting to investigate how the effect of contextual features (e.g. age, gender, and relationship) would change if the age window for the participants were widened. Also, we focus solely on textual and contextual information, rather than media (e.g., images, videos, links), which has been the subject of inquiry of prior work in the HCI community [3, 72]. In future, our aim is to consider multi-modal approaches for sexual risk detection that includes both textual content and media.

While our results are promising for detecting unsafe sexual conversations experienced by youth, we faced the challenge of imbalance in our dataset, having relatively smaller unsafe interactions than safe ones. For creating a balanced dataset for the unsafe sexual conversation classifier we used under-sampling to reduce the number of safe conversations. The main issue with under-sampling is the possibility of losing informative instances from the majority class while deleting the instances. To make sure that the sampled examples were diverse enough, we manually checked the remaining samples. Yet, before our models are deployed in large, user-based platforms, more unsafe training samples are needed to reduce false-positives. To address these limitations, we are in the process

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

of collecting more data. As a future work, we also plan to have youth evaluate the quality of our sexual risk classifiers by designing and deploying a web-based risk detection system where they can upload their Instagram data for the system to identify risky content. This will complete the HCML loop of having users direct feedback on the performance of our algorithms, so that they can be further refined for future real-world use and impact. A major contribution of the work is the dataset and we call for more researchers to join us to work on this topic using the dataset [17]. But, due to the sensitive nature of the data (e.g., the risk experiences of youth) and because social media data cannot truly be anonymized, we chose to prioritize participant privacy [2, 57] and confidentiality over public replicability of our results. However, we will mitigate the concern of replicability in the following ways: 1) We are open to collaborate with others in the research community by sharing the dataset under a licensing agreement that ensures participant protection, 2) We are exploring the feasibility of anonymizing portions of the dataset to share publicly. We also created an open-source community [17] for collaboration for the purpose of advancing the field and replicability of research. This open-source community would bring together a diverse group of people for the goal of youth online safety to contribute to advance the state-of-the-art in algorithmic risk detection.

7 CONCLUSION

The core contribution of this work is that our findings are grounded in the voices of youth who experienced online sexual risks and were brave enough to share these experiences with us. To the best of our knowledge, this is the first work that analyzes machine learning approaches on private social media conversations of youth to detect unsafe sexual conversations. In addition, this work highlights the importance of contextual and implicit features on identifications of unsafe sexual conversations and provides a good indication of how different methods and features perform when addressing this problem. Given the wealth of data we have collected, but have yet to analyze, we welcome other HCI and ML researchers to join us in our efforts to use HCML as a proactive way to protect and empower youth online.

ACKNOWLEDGMENTS

This research is supported in part by the U.S. National Science Foundation under grants #IIP-1827700, #IIS-1844881, #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 towards our research.

REFERENCES

[1] 2020. *National Center for Missing Exploited Children.* https://www.missingkids.org/footer/media/keyfacts

[2] Zainab Agha, Neeraj Chatlani, Afsaneh Razi, and Pamela Wisniewski. 2020. Towards Conducting Responsible Research with Teens and Parents Regarding Online Risks. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/3334480.3383073

[3] 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. https://doi.org/10.1145/3491102.3501969

[4] Ashwaq Alsoubai, Xavier V. Caddle, Ryan Doherty, Alexandra Taylor Koehler, Estefania Sanchez, Munmun De Choudhury, and Pamela J. Wisniewski. 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* (Virtual Event, Taiwan) *(CSCW'22 Companion)*. Association for Computing Machinery, New York, NY, USA, 197–200. https://doi.org/10.1145/3500868.3559710

[5] 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

[6] Nazanin Andalibi, Oliver L. Haimson, Munmun De Choudhury, and Andrea Forte. 2016. Understanding Social Media Disclosures of Sexual Abuse Through the Lenses of Support Seeking and Anonymity. In *CHI '16*. ACM Press, 3906–3918. https://doi.org/10.1145/2858036.2858096

[7] 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/

[8] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (2020), 82–115.

[9] M. Ashcroft, L. Kaati, and M. Meyer. 2015. A Step Towards Detecting Online Grooming – Identifying Adults Pretending to be Children. In *2015 European Intelligence and Security Informatics Conference*. 98–104. https://doi.org/10.1109/EISIC.2015.41

[10] Eric PS Baumer. 2017. Toward human-centered algorithm design. *Big Data & Society* 4, 2 (Dec. 2017), 2053951717718854. https://doi.org/10.1177/2053951717718854

[11] Susanne E. Baumgartner, Patti M. Valkenburg, and Jochen Peter. 2010. Unwanted online sexual solicitation and risky sexual online behavior across the lifespan. *Journal of Applied Developmental Psychology* 31, 6 (Nov. 2010), 439–447. https://doi.org/10.1016/j.appdev.2010.07.005

[12] Markus Bayer, Marc-André Kaufhold, and Christian Reuter. 2021. A survey on data augmentation for text classification. *arXiv preprint arXiv:2107.03158* (2021).

[13] 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.

[14] Michael Brenner, Jeannie Suk Gersen, Michael Haley, Matthew Lin, Amil Merchant, Richard Jagdishwar Millett, Suproteem K Sarkar, and Drew Wegner. 2020. Constitutional Dimensions of Predictive Algorithms in Criminal Justice. *Harv. CR-CLL Rev.* 55 (2020), 267.

[15] Laura Jayne Broome, Cristina Izura, and Jason Davies. 2020. A psycho-linguistic profile of online grooming conversations: A comparative study of prison and police staff considerations. *Child Abuse Neglect* 109 (2020), 104647. https://doi.org/10.1016/j.chiabu.2020.104647

[16] Amy Bruckman. 2002. Studying the amateur artist: A perspective on disguising data collected in human subjects research on the Internet. *Ethics and Information Technology* 4, 3 (2002), 217–231.

[17] Xavier V Caddle, Afsaneh Razi, Seunghyun Kim, Shiza Ali, Temi Popo, Gianluca Stringhini, Munmun De Choudhury, and Pamela J Wisniewski. 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*. 315–318.

[18] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16 (2002), 321–357.

[19] Maryam Daniali and Edward Kim. 2022. Perception Over Time: Temporal Dynamics for Robust Image Understanding. *arXiv preprint arXiv:2203.06254* (2022).

[20] Munmun De Choudhury and Sushovan De. 2014. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *8th Intl AAAI conference on weblogs and social media*.

[21] Patricia De Santisteban and Manuel Gámez-Guadix. 2018. Prevalence and risk factors among minors for online sexual solicitations and interactions with adults. *J. Sex Research* 55, 7 (2018), 939–950.

[22] Prema Dev, Jessica Medina, Zainab Agha, Munmun De Choudhury, Afsaneh Razi, and Pamela J Wisniewski. 2022. From Ignoring Strangers' Solicitations to Mutual Sexting with Friends: Understanding Youth's Online Sexual Risks in Instagram Private Conversations. In *Companion Publication of the 2022 Conference on Computer Supported Cooperative Work and Social Computing*. 94–97.

[23] Allyson L. Dir, Ayca Coskunpinar, Jennifer L. Steiner, and Melissa A. Cyders. 2013. Understanding Differences in Sexting Behaviors Across Gender, Relationship Status, and Sexual Identity, and the Role of Expectancies in Sexting. *Cyberpsychology, Behavior, and Social Networking* 16, 8 (May 2013), 568–574. https://doi.org/10.1089/cyber.2012.0545

[24] Allyson L Dir, Ayca Coskunpinar, Jennifer L Steiner, and Melissa A Cyders. 2013. Understanding differences in sexting behaviors across gender, relationship status, and sexual identity, and the role of expectancies in sexting. *Cyberpsychology, Behavior, and Social Networking* 16, 8 (2013), 568–574.

[25] Michelle Drouin and Elizabeth Tobin. 2014. Unwanted but consensual sexting among young adults: Relations with attachment and sexual motivations. *Computers in Human Behavior* 31 (2014), 412–418.

[26] Muhammad Ali Fauzi and Patrick Bours. 2020. Ensemble Method for Sexual Predators Identification in Online Chats. In *8th International Workshop on Biometrics and Forensics*. 1–6. https://doi.org/10.1109/IWBF49977.2020.9107945

[27] Manas Gaur, Keyur Faldu, and Amit Sheth. 2021. Semantics of the black-box: Can knowledge graphs help make deep learning systems more interpretable and explainable? *IEEE Internet Computing* 25, 1 (2021), 51–59.

[28] Arijit Ghosh Chowdhury, Ramit Sawhney, Puneet Mathur, Debanjan Mahata, and Rajiv Ratn Shah. 2019. Speak up, Fight Back! Detection of Social Media Disclosures of Sexual Harassment. In *NNACL*. Association for Computational Linguistics, Minneapolis, Minnesota, 136–146. https://doi.org/10.18653/v1/N19-3018

[29] Ana M Giménez Gualdo, Simon C Hunter, Kevin Durkin, Pilar Arnaiz, and Javier J Maquilón. 2015. The emotional impact of cyberbullying: Differences in perceptions and experiences as a function of role. *Computers & Education* 82 (2015), 228–235.

[30] Eszter Hargittai, Jason Schultz, John Palfrey, et al. 2011. Why parents help their children lie to Facebook about age: Unintended consequences of the 'Children's Online Privacy Protection Act'. *First Monday* (2011).

[31] Heidi Hartikainen, Afsaneh Razi, and Pamela Wisniewski. 2021. Safe Sexting: The Advice and Support Adolescents Receive from Peers Regarding Online Sexual Risks. *Proc. ACM-HCI* 5, CSCW1, Article 42 (April 2021), 31 pages. https://doi.org/10.1145/3449116

[32] Heidi Hartikainen, Afsaneh Razi, and Pamela Wisniewski. 2021. '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 Conference on Computer Supported Cooperative Work and Social Computing*. 67–71.

[33] 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 International AAAI Conference on Web and Social Media*, Vol. 14. 250–259.

[34] Laurie Collier Hillstrom. 2018. *The# metoo movement*. ABC-CLIO.

[35] Michael Hind, Sameep Mehta, Aleksandra Mojsilovic, Ravi Nair, Karthikeyan Natesan Ramamurthy, Alexandra Olteanu, and Kush R Varshney. 2018. Increasing trust in ai services through supplier's declarations of conformity. *arXiv preprint arXiv:1808.07261* 18 (2018), 2813–2869.

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

[37] 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.

[38] 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 *Proc CSCW (CSCW '15)*. ACM, New York, NY, USA, 583–599. https://doi.org/10.1145/2675133.2675287

[39] Sweta Karlekar and Mohit Bansal. 2018. SafeCity: Understanding Diverse Forms of Sexual Harassment Personal Stories. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 2805–2811. https://doi.org/10.18653/v1/D18-1303

[40] Seunghyun Kim, Afsaneh Razi, Gianluca Stringhini, Pamela Wisniewski, and Munmun De Choudhury. 2021. You Don't Know How I Feel: Insider-Outsider Perspective Gaps in Cyberbullying Risk Detection. In *Proc. Intl AAAI Conference on Web and Social Media*.

[41] Seunghyun Kim, Afsaneh Razi, Gianluca Stringhini, Pamela J Wisniewski, and Munmun De Choudhury. 2021. A Human-Centered Systematic Literature Review of Cyberbullying Detection Algorithms. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–34.

[42] Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *EMNLP*.

[43] Sosuke Kobayashi. 2018. Contextual augmentation: Data augmentation by words with paradigmatic relations. *arXiv preprint arXiv:1805.06201* (2018).

[44] Kat Krol, Matthew Moroz, and M. Angela Sasse. 2012. Don't work. Can't work? Why it's time to rethink security warnings. In *Intl Conf on Risks and Security of Internet and Systems (CRiSIS)*. 1–8. https://doi.org/10.1109/CRISIS.2012.6378951

[45] Amanda Lenhart, Monica Anderson, and Aaron Smith. 2015. Teens, Technology and Romantic Relationships | Pew Research Center. http://www.pewinternet.org/2015/10/01/teens-technology-and-romantic-relationships/

[46] Hyeontaek Lim, David G Andersen, and Michael Kaminsky. 2019. 3lc: Lightweight and effective traffic compression for distributed machine learning. *Proceedings of Machine Learning and Systems* 1 (2019), 53–64.

[47] Yingchi Liu, Quanzhi Li, Xiaozhong Liu, Qiong Zhang, and Luo Si. 2019. Sexual Harassment Story Classification and Key Information Identification *(CIKM '19)*. ACM, Beijing, China, 2385–2388. https://doi.org/10.1145/3357384.3358146

[48] Nuria Lorenzo-Dus, Anina Kinzel, and Matteo Di Cristofaro. 2020. The communicative modus operandi of online child sexual groomers: Recurring patterns in their language use. *Journal of Pragmatics* 155 (2020), 15–27. https://doi.org/10.1016/j.pragma.2019.09.010

[49] Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proc. ACL: Sys Dem.* 55–60.

[50] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635* (2019).
[51] Rachel O'Connell. 2003. A typology of child cybersexploitation and online grooming practices. *Cyberspace Research Unit, University of Central Lancashire* (2003).
[52] James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates* 71, 2001 (2001), 2001.
[53] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe Global Vectors for Word Representation. , 1532–1543 pages. http://www.aclweb.org/anthology/D14-1162
[54] Ralph M Perhac Jr. 1996. Defining risk: Normative considerations. *Human and Ecological Risk Assessment* 2, 2 (1996), 381–392.
[55] Arun Rai. 2020. Explainable AI: From black box to glass box. *Journal of the Academy of Marketing Science* 48, 1 (2020), 137–141.
[56] Gonzalo Ramos, Jina Suh, Soroush Ghorashi, Christopher Meek, Richard Banks, Saleema Amershi, Rebecca Fiebrink, Alison Smith-Renner, and Gagan Bansal. 2019. Emerging Perspectives in Human-Centered Machine Learning. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–8.
[57] Afsaneh Razi, Zainab Agha, Neeraj Chatlani, and Pamela Wisniewski. 2020. Privacy Challenges for Adolescents as a Vulnerable Population. In *Networked Privacy Workshop of the 2020 CHI Conference on Human Factors in Computing Systems*.
[58] 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 *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI EA '22)*. Association for Computing Machinery, New York, NY, USA, Article 39, 9 pages. https://doi.org/10.1145/3491101.3503569
[59] 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 *Proc 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. ACM, Honolulu, HI, USA, 1–13. https://doi.org/10.1145/3313831.3376400
[60] Afsaneh Razi, Seunghyun Kim, Ashwaq Soubai, Gianluca Stringhini, Thamar Solorio, Munmun De Choudhury, and Pamela 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
[61] Joseph Reagle. 2022. Disguising Reddit sources and the efficacy of ethical research. *Ethics and Information Technology* 24, 3 (2022), 1–11.
[62] Joseph Reagle and Manas Gaur. 2022. Spinning words as disguise: Shady services for ethical research? *First Monday* (2022).
[63] Lauren Reed, Margaret Boyer, Haley Meskunas, Richard Tolman, and L Ward. 2020. How do adolescents experience sexting in dating relationships? Motivations to sext and responses to sexting requests from dating partners. *Children & Youth Services Rev* 109 (2020).
[64] Tatiana R. Ringenberg, Kanishka Misra, and Julia Taylor Rayz. 2019. Not So Cute but Fuzzy: Estimating Risk of Sexual Predation in Online Conversations. In *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*. 2946–2951. https://doi.org/10.1109/SMC.2019.8914528 ISSN: 1062-922X.
[65] Dominik Sacha, Michael Sedlmair, Leishi Zhang, John Aldo Lee, Daniel Weiskopf, Stephen North, and Daniel Keim. 2016. Human-centered machine learning through interactive visualization. ESANN, Bruges, Belgium, 641–646. https://www.elen.ucl.ac.be/Proceedings/esann/esannpdf/es2016-166.pdf
[66] Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. 2008. *Introduction to information retrieval*. Cambridge University Press.
[67] Kathryn C. Seigfried-Spellar, Marcus K. Rogers, Julia T. Rayz, Shih-Feng Yang, Kanishka Misra, and Tatiana Ringenberg. 2019. Chat Analysis Triage Tool: Differentiating contact-driven vs. fantasy-driven child sex offenders. *Forensic Science International* (Feb. 2019). https://doi.org/10.1016/j.forsciint.2019.02.028
[68] Donald Sharpe. 2015. Chi-square test is statistically significant: Now what? *Practical Assessment, Research, and Evaluation* 20, 1 (2015), 8.
[69] Zipora Shechtman, David L Vogel, Haley A Strass, and Patrick J Heath. 2018. Stigma in help-seeking: the case of adolescents. *British Journal of Guidance & Counselling* 46, 1 (2018), 104–119.
[70] Svetlana Smirnova, Sonia Livingstone, and Mariya Stoilova. 2021. Understanding of user needs and problems: a rapid evidence review of age assurance and parental controls. (2021).
[71] Ashima Suvarna, Grusha Bhalla, Shailender Kumar, and Ashi Bhardwaj. 2020. Identifying Victim Blaming Language in Discussions about Sexual Assaults on Twitter. In *Intl Conf on Social Media and Society (SMSociety'20)*. ACM, Toronto, ON, Canada, 156–163. https://doi.org/10.1145/3400806.3400825

[72] Muhammad Uzair Tariq, Afsaneh Razi, Karla Badillo-Urquiola, and Pamela Wisniewski. 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 (Lecture Notes in Computer Science)*, Masaaki Kurosu (Ed.). Springer International Publishing, Cham, 90–108. https://doi.org/10.1007/978-3-030-22636-7_6

[73] Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology* 29, 1 (2010), 24–54.

[74] David Van Bruwaene, Qianjia Huang, and Diana Inkpen. 2020. A multi-platform dataset for detecting cyberbullying in social media. *Language Resources and Evaluation* 54, 4 (2020), 851–874.

[75] Joris Van Ouytsel, Michel Walrave, Koen Ponnet, and Wannes Heirman. 2015. The association between adolescent sexting, psychosocial difficulties, and risk behavior: Integrative review. *The Journal of School Nursing* 31, 1 (2015), 54–69.

[76] Jennifer Wortman Vaughan and Hanna Wallach. 2020. A human-centered agenda for intelligible machine learning. *Machines We Trust: Getting Along with Artificial Intelligence* (2020).

[77] Jessica Vitak and Jinyoung Kim. 2014. " You can't block people offline" examining how facebook's affordances shape the disclosure process. In *In Proc 17th ACM CSCW*. 461–474.

[78] Helen Whittle, Catherine Hamilton-Giachritsis, Anthony Beech, and Guy Collings. 2013. A review of online grooming: Characteristics and concerns. *Aggression and Violent Behavior* 18, 1 (Jan. 2013), 62–70. https://doi.org/10.1016/j.avb.2012.09.003

[79] 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 *Proc 2016 CHI (CHI '16)*. ACM, New York, NY, USA, 3919–3930. https://doi.org/10.1145/2858036.2858317 San Jose, California, USA.

[80] Michele L Ybarra and Kimberly J Mitchell. 2008. How risky are social networking sites? A comparison of places online where youth sexual solicitation and harassment occurs. *Pediatrics* 121, 2 (2008).

[81] Fabio Massimo Zanzotto. 2019. Human-in-the-loop Artificial Intelligence. *Journal of Artificial Intelligence Research* 64 (2019), 243–252.

A PRE-DEFINED RISK TYPES AND LEVELS

Drawing on a set of pre-defined risk types derived in a domain-driven manner from existing Instagram reporting feature risk categories [15], we explained to participants that unsafe or uncomfortable interactions may include but were not limited to:

- **Nudity/porn:** Photos or videos of nude or partially nude people or person.
- **Sexual messages or Solicitations:** Sending or receiving a sexual message ("Sexting") – being asked to send a sexual message, revealing, or naked photo.
- **Harassment:** Messages that contain credible threats, aim to degrade or shame someone, contain personal information to blackmail or harass someone, or threaten to post nude photos of someone.
- **Hate speech:** Messages that encourage violence or attack anyone based on who they are; specific threats of physical harm, theft, or vandalism.
- **Violence/Threat of violence:** Messages, photos, or videos of extreme violence, or that encourage violence or attacks anyone based on their religious, ethnic, or sexual background.
- **Sale or promotion of illegal activities:** Messages promoting the use, or distributing illegal material such as drugs.
- **Self-injury:** Messages promoting self-injury, which includes suicidal thoughts, cutting, and/or eating disorders.
- **Other:** Other situations that could potentially lead to emotional or physical harm.

We then grounded risk levels based in the existing adolescent online risk literature [79] which operationalized the risk level for youth for how much it is likely to cause emotional or physical harm to them or others:

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

- **Low Risk** comprised messages that made the participant uncomfortable but were unlikely to cause emotional or physical harm.
- **Medium Risk** included messaging which if continued/escalated, would have been likely to cause emotional/physical harm.
- **High Risk** comprised messages that were deemed dangerous and caused emotional or physical harm to the participant.

Received January 2022; revised July 2022; accepted November 2022