A Human-Centered Systematic Literature Review of the Computational Approaches for Online Sexual Risk Detection

AFSANEH RAZI, University of Central Florida, U.S.A
SEUNGHYUN KIM, Georgia Institute of Technology, U.S.A
ASHWAQ SOUBAI, University of Central Florida, U.S.A
GIANLUCA STRINGHINI, Boston University, U.S.A
THAMAR SOLORIO, University of Houston, U.S.A
MUNMUN DE CHODHURY, Georgia Institute of Technology, U.S.A
PAMELA WISNIEWSKI, University of Central Florida, U.S.A

In the era of big data and artificial intelligence, online risk detection has become a popular research topic. From detecting online harassment to the sexual predation of youth, the state-of-the-art in computational risk detection has the potential to protect particularly vulnerable populations from online victimization. Yet, this is a high-risk, high-reward endeavor that requires a systematic and human-centered approach to synthesize disparate bodies of research across different application domains, so that we can identify best practices, potential gaps, and set a strategic research agenda for leveraging these approaches in a way that betters society. Therefore, we conducted a comprehensive literature review to analyze 73 peer-reviewed articles on computational approaches utilizing text or meta-data/multimedia for online sexual risk detection. We identified sexual grooming (75%), sex trafficking (12%), and sexual harassment and/or abuse (12%) as the three types of sexual risk detection present in the extant literature. Furthermore, we found that the majority (93%) of this work has focused on identifying sexual predators after-the-fact, rather than taking more nuanced approaches to identify potential victims and problematic patterns that could be used to prevent victimization before it occurs. Many studies rely on public datasets (82%) and third-party annotators (33%) to establish ground truth and train their algorithms. Finally, the majority of this work (78%) mostly focused on algorithmic performance evaluation of their model and rarely (4%) evaluate these systems with real users. Thus, we urge computational risk detection researchers to integrate more human-centered approaches to both developing and evaluating sexual risk detection algorithms to ensure the broader societal impacts of this important work.

CCS Concepts: • **Human-centered computing** → **Empirical studies in collaborative and social computing**.

Additional Key Words and Phrases: Sexual Risk Detection, Human-Centered Machine Learning, Online Risks, Artificial Intelligence, Literature Review, Social Media

Authors' addresses: Afsaneh Razi, afsaneh.razi@knights.ucf.edu, University of Central Florida, 4000, Orlando, Florida, U.S.A, 32816; Seunghyun Kim, Georgia Institute of Technology, 30318, Atlanta, Georgia, U.S.A, seunghyun.kim@gatech.edu; Ashwaq Soubai, University of Central Florida, 4000, Orlando, Florida, U.S.A, atalsoubai@Knights.ucf.edu; Gianluca Stringhini, Boston University, 02215, Boston, Massachusetts, U.S.A, gian@bu.edu; Thamar Solorio, University of Houston, 584, Houston, Texas, U.S.A, tsolorio@uh.edu; Munmun De Chodhury, Georgia Institute of Technology, 30318, Atlanta, Georgia, U.S.A, munmund@gatech.edu; Pamela Wisniewski, University of Central Florida, 4000, Orlando, Florida, U.S.A, pamwis@ucf.edu.

2573-0142/2021/10-ART465 $15.00
https://doi.org/10.1145/3479609

ACM Reference Format:
Afsaneh Razi, Seunghyun Kim, Ashwaq Soubai, Gianluca Stringhini, Thamar Solorio, Munmun De Chodhury, 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 (October 2021), 38 pages. https://doi.org/10.1145/3479609

1 INTRODUCTION

The internet has the potential to facilitate new friendships and romantic partnerships [134], but it has also been used to commit sexually-based harm [56, 144]. As such, technology-facilitated sexual violence, which manifests in non-consensual or unwanted sexting, sexual grooming, sex trafficking, and/or exploitation or abuse, has become a prevalent concern among internet users [56]. For instance, the National Center for Missing and Exploited Children (NCMEC) [2] received more than 16.9 million child sexual exploitation reports in 2019, which included online child grooming, sextortion, and the engagement of children in sexual activity via the internet. Online grooming behaviors can also lead to sex trafficking [12], which is "the recruitment, harboring, transportation, provision, or obtaining of a person for the purpose of a commercial sex act" [3]. With the rise of the #MeToo movement [58], sexual harassment and/or abuse has also become a popular research topic within the SIGCHI and Computer-Supported Cooperative Work and Social Computing (CSCW) research communities, ranging from in-depth qualitative accounts of sexual victimization [8, 52] to computational approaches for sexual risk detection [88, 147].

In 2018, internet-based sexual violence was thrust to the forefront of U.S. political legislation when the Fight Online Sex Trafficking Act (FOSTA) and Stop Enabling Sex Traffickers Act (SESTA) bills [132] were signed and immediately put into effect. Together, these bills, for the first time in history, made online platforms accountable for sex trafficking facilitated via their platforms. Given the political landscape and the prevalence of internet-based sexual violence, the relevance and applicability of machine learning (ML) approaches for online sexual risk detection has become of critical importance when considering solutions for solving this dire societal problem. As such, researchers have begun trying to synthesize the state-of-the-art in computational ML for various online sexual risks [70, 90, 97, 149]. For instance, Tariq et al. [149] surveyed Computer Vision (CV) approaches for detecting skin dominance and nudity in digital images and videos for the purpose of combating adolescent sexting. They identified the need for more human-centeredness in the design and development of nudity detection algorithms to make them applicable for real-world deployment. We build upon this prior literature by conducting a human-centered systematic review of the computational approaches for online sexual risk detection within text-based and multi-modal data (i.e., text with images, videos, and/or meta data). We define computational risk detection as ML and other automated approaches that predict risk-related behavior [104].

One may ask why we need human-centeredness in computational approaches for online sexual risk detection. Artificial Intelligence (AI) systems are used to make decisions across various human domains, and they impact people's lives in high-stake situations such as criminal justice [62], child welfare [139], and healthcare [32]. Past research has expressed concerns about these systems failing to account for limitations and/or uncertainties inherent in their predictions that may result in negative impacts to people's lives [154]. As such, "human-centered machine learning" (HCML) is an emerging sub-field of Computer Science that leverages computational AI expertise and human knowledge from the social sciences to ensure that ML approaches address societal needs and do no harm. For instance, human-centeredness provides insight into the potential pitfalls and ethical considerations inherent in using technology to solve human problems that a purely computational lens lacks, providing a better understanding of how these models will perform in the wild and the potential impact these technologies will have on end users [71]. Moreover, Human-Centered

Design (HCD) enables researchers to incorporate the perspectives of various stakeholders, helping them to construct a robust algorithm [128, 138]; in our context, leveraging such a human-centered lens can facilitate reaching the goal of detecting, mitigating, and preventing online sexual risks.

Traditional frameworks of computational ML risk detection focus on the 1) **Data** used to train the algorithms, 2) **Models** or algorithmic approaches for risk detection, and 3) **Evaluation** metrics for assessing how these models perform. Yet, we used a human-centered lens to conduct a systematic literature review of computational approaches for online sexual risk detection that led us to ask more nuanced research questions:

- **RQ1 (Data):** *Are the datasets ecologically valid for detecting the targeted risk for the desired user population?*
- **RQ2 (Models):** *Are the algorithmic models grounded in human theory, understanding, and knowledge?*
- **RQ3 (Evaluation):** *How well do the algorithms perform, both computationally and in meeting end users' needs?*
- **RQ4 (Application):** *What system artifacts were developed, and what were the outcomes when deployed in real-world settings?*

To answer these research questions, we conducted a systematic literature review which analyzed 73 peer-reviewed papers published between 2007 and 2020. We performed a comprehensive literature search to identify any computational approaches for online sexual risk detection within text-based and multi-modal data. In our literature review, we broadly considered all types of online sexual risks that may result in mental or physical harm, including sexual violence/abuse, sexual harassment, sexual grooming, and sex trafficking. We qualitatively coded these articles using a human-centered lens that assessed the ecological validity of the data being used to train the algorithms, the algorithmic approaches being used, the metrics for which to assess the quality of these models, and whether and how these models were deployed in real-world settings.

Overall, we found that most papers proposed algorithms for detecting sexual predators (75%) after the sexual violence occurred (93%) using public datasets (82%). These findings imply that there is a need for approaches that help prevent victimization and to detect other types of sexual risks, such as sexting. We identified that most (52%) approaches relied on datasets that are not representative of end users, and are annotated by third parties without adequate background on the subject (31%). The takeaway from this finding is that it is crucial to have realistic datasets with high-quality annotations for training models that take into account the perspectives of the individuals who experienced the risk as well as experts and clinicians to do not exclusively rely on researchers and volunteers. We found that most (60%) models do a binary classification of users assessing whether they are a predator or not; there is a need for approaches that take into account patterns and changes at the conversation level so that detecting the risky content will be more effective and meaningful. The results of our review for evaluation methods illustrate that the main (81%) focus is on computational evaluation of the method and that there are only a few (8%) user studies to evaluate the developed technology. In addition, most (93%) of the studies proposed algorithms but did not integrate these algorithms in a real system that could be used by stakeholders to identify and mitigate sexual risks. Our research makes the following novel contributions to the CSCW, HCML, and ML research communities:

- A conceptual framework for systematically reviewing computation risk detection literature using a human-centered lens (Figure 1)
- An in-depth synthesis of the current state-of-the-art and trends in computational approaches for online sexual risk detection

- Identification of the potential gaps within the existing literature and recommendations for a research agenda that would advance beyond the state-of-the-art within this research domain

In the following section, we describe the existing efforts that have been taken in the area of computational sexual risk detection. Then, we introduce our human-centered framework for conducting our computational literature review and our methods in which we instantiated this framework to qualitatively analyze the literature.

2 BACKGROUND

Prior reviews of computational approaches for sexual risk detection were typically couched more broadly in the examination and detection of online abuse and cyber-aggression. For instance, a review by Mishra et al. [100] presented the computational approaches for detecting online abuse, including racism, sexism, personal attacks, toxicity, and harassment. They took a coarse-grained definition of abuse as "any expression that is meant to denigrate or offend a particular person or group", in their review of the literature. Similarly, Mladenovic et al. [103] reviewed papers focused on detecting cyber-aggression, cyberbullying, and cyber-grooming. The review by Mladenovic et al. [103] took a generalized view in reviewing papers that included any risks (including sexual) within aggressor–victim relationships in online platforms. In both reviews [100, 103], they found those deep learning models, such as Convolutional Neural Networks (CNNs), achieved better accuracy than traditional machine learning models. Mladenovic et al. [103] also found that the most useful features are word and character embeddings. They also mention that English is the pre-eminent language studied from the perspective of all the three reviewed risks and there is a gap of having datasets in different languages. These researchers made notable contributions by synthesizing the literature and identifying potential research gaps using a primarily computational lens.

From a human-centered perspective, we argue that sexual risks are a unique form of violence and that reviewing literature for different risk phenomenons (e.g., cyberbullying vs. cyber-grooming), under the assumption they are similar, may lead to false conclusions or unintended and/or negative consequences for sexual violence victims. Different risk types have different characteristics of victimization that should be taken into consideration; for example, cyberbullying and cyber-grooming are distinct risk types, even though they share a generalized definition as an attack directed to harm a victim [103]. Cyberbullying includes an intentional and aggressive act against someone or a group of people while cyber-grooming happens between a sexual predator and victim to gain the victim's trust for the purpose of sexual abuse [103]. Each of these risks deserves careful attention and exploration. In recent years, researchers made efforts to incorporate unique characteristics of a specific risk type in their models[34]. Therefore, we reviewed papers specific to sexual risk detection, which includes sexual grooming, sex trafficking, and sexual harassment and/or abuse.

A couple of reviews have been conducted on online sexual risk detection in the specific context of sexual grooming and identification of child sexual predators. These reviews coincided with the 2012 Sexual Predator Identification competition ran by PAN[1] using the PAN-12 dataset. This competition aimed to provide researchers with an initial benchmark for comparing different methods of detecting cyberpeodophiles or sexual groomers by using PAN-12, a large dataset of chat logs between convicted sex offenders and volunteers posing as children (created by Perverted Justice [67]). Inches and Crestani [67] published a review of approaches that were taken during the competition. In their review, they discussed the submissions' pre-filtering techniques, features, the classification models, and evaluated the computational approaches submitted by 16 teams. The top five teams who were able to identify sexual predators attained a higher accuracy using Support Vector Machine (SVM) algorithms. This research provided a framework for benchmarking the

[1] A benchmarking activity on uncovering plagiarism, authorship and social software misuse http://pan.webis.de

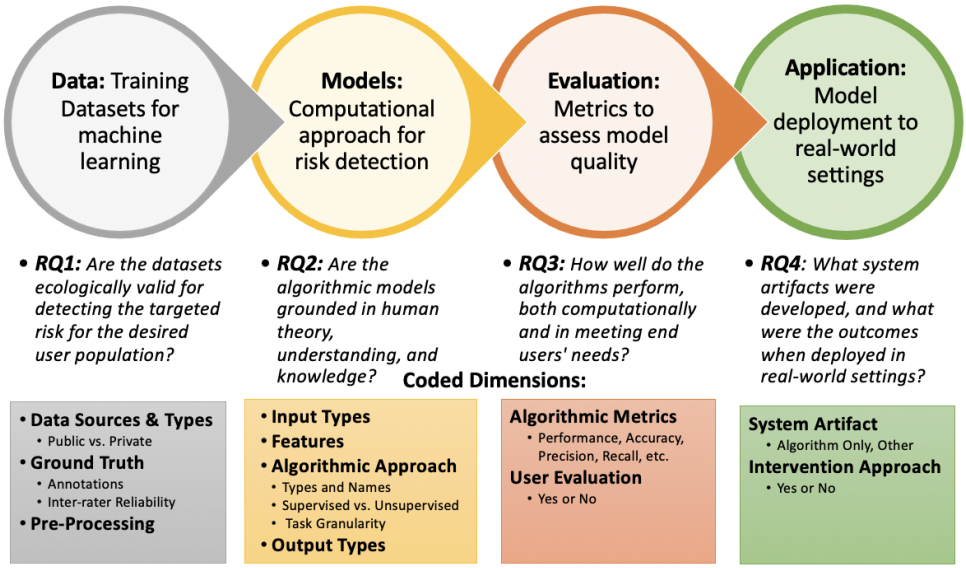

Fig. 1. Human-Centered Lens for Computational Risk Detection Systematic Literature Reviews

computational approaches for sexual predator identification from a technical standpoint. Ngejane et al.[108] continued the effort by reviewing 10 additional papers that used the PAN-12 dataset that were published after Inches' and Crestani's [67] review and found that most of the reviewed approaches used supervised models; among these supervised models, they confirmed that SVM yielded the highest accuracy (98%) in detecting child sexual predators in the PAN-12 data.

A common theme among these prior syntheses of the computational literature for both generalized and/or sexual risk detection is that they focused predominantly on the algorithms themselves in terms of the standard ML metrics for benchmarking performance. Although quantitative measures are necessary for evaluating the algorithms' performance, quantitative measures are not sufficient in determining if the models perform well from the point of view of stakeholders and end users. Hence, these computationally-focused reviews suffer a disconnect between the impact on the functionality of the algorithms and the social interpretations needed to assess quality in the broader sense of the problem context [17]. As Baumer [17] suggests, these disconnects can be addressed by incorporating human-centeredness to evaluate the approaches with deeper interpretations behind algorithmic inferences that impact real-world use [17]. Unlike the previous works, our review leverages this human-centered lens to synthesize the computational literature on detecting online sexual risks. Sexual victimization is a social issue that requires the integration of personal, social, and cultural aspects for designing and developing intelligent systems that understand this socially complicated issue. Thus, our work aims to bridge the gap between social needs and computational perspectives by using a human-centered lens to synthesize the computational risk detection literature in this domain. In the next section, we introduce our human-centered lens for reviewing computational risk detection in the context of sexual risk detection.

3 APPLYING A HUMAN-CENTERED LENS TO COMPUTATIONAL RISK DETECTION

We present four main components of computational ML systems (i.e., Data, Models, Evaluation, and Application) and demonstrate how a human-centered lens can be applied for analyzing computational risk detection research (Figure 1). Traditionally, ML researchers focus on data, models, and computational evaluations of these models. These are the main components of ML which every systematic literature review should take into account; yet, these components should also consider the human-context in which these models intend to be deployed. To do this, we synthesized relevant bodies of work across the fields of Human-Computer Interaction (HCI), HCML, and ML to create a conceptual framework in which to conduct our literature review. This framework is one of the contributions of this paper and also served as a theoretical lens for grounding the analyses of the reviewed papers within the domain of online sexual risk detection. While we apply this framework to the specific context of online sexual risk detection, it can also be generalized to other forms of computation risk detection that could benefit from a human-centered perspective.

The ubiquitous use of systems to produce risk predictions have consequences that impact people's lives [111]; ML/AI systems have been used in decision making systems in various contexts [6], such as child welfare [139], that affect people's lives in profound ways. Recently, researchers across multidisciplinary fields, including HCI, ML, public policy, and humanities have come together to address the gaps between computational AI systems and societal needs. This had led to the paradigm shift toward HCML, which tries to address a system's inability to improvise according to context and human characteristics such as perceptions, emotions, intentions, and social contexts [18, 158] to keep humans at the center of the design process by taking into account stakeholders needs. Yet, researchers have varied on a formal definition for human-centered computing; Kling and Star expressed that "there is no simple recipe for the design or use of human-centered computing" [78]. Chancellor et al. [31] argued that HCML is a growing interdisciplinary field that there are no formal definitions for, so they sought to understand the definition of "human" in regard to HCML for mental health. Similarly, we apply a human-centered lens broadly for reviewing computational approaches for online sexual risk detection.

Importantly, there are key differences that distinguish computational risk detection from general computational ML predictions that need to be considered when reviewing the literature. Risk is a subjective concept, which makes it difficult to operationalize [120]. Thus, it is important to look at the specific problem and the context around which risks occur. As the goal of detecting risk is typically to protect people from harm, inaccurate risk detection could be potentially life-altering, creating real-world ramifications for both potential victims and alleged predators. When applying ML to risk detection, false positives or/and negatives may have adverse consequences for users. As such, computational risk detection is a high-risk, high-reward research problem, which requires extra scrutiny. In Figure 1, we reiterate our high-level research questions and map these questions to dimensions of the literature that coded for in our systematic review of the sexual risk detection literature. In the sections below, we describe each component of our human-centered framework in more detail.

3.1 Considerations for Data Using a Human-Centered Lens

Data is defined as sets of instances for building or evaluating ML algorithms, and label is a value or category assigned to each data instance and served as the target for the algorithm [104]. Given the fact that data is the foundation of algorithmic development, it is important that the data matches the real world users' context [53]. Knowing whether the data is the true representation of user behavior is important, so computational risk detection research should consider the motivations behind how the dataset was created, as well as the characteristics of the data itself. The nature of the risks that

happen on various platforms is distinct, as each online platform has its own characteristics and affordances [93]. For example, the nature of the risks in public posts on Twitter may be different than the nature of private conversations on Facebook. Researchers also need to review the mechanisms or procedures of the data collection and ground truth annotations to examine if there were clear explanations behind the process as well as steps to validate such procedures. Researchers should also analyze the types of data included in the dataset. When it comes to detecting risks, important indicators, such as the context of the relationship [127] or time of communication [89] may be indicative of risk. Since risks are context-driven [120], having multi-modal data points could help improve model performance.

Other data considerations may include privacy policies, consent processes (if data is collected from human subjects), ethical considerations, and potential sampling bias. For instance, the frequency and the nature of risks that occur in private spaces are different than the risks that happen in public spaces [161]. Having transparency about the data collection process would help users and researchers to identify any types of biases or data privacy issues, to ensure correct use and distribution of the data. For instance, if researchers are using public datasets to detect risks, then the nature of risks will be limited to these public contexts. Additionally, how the data is labeled for the ground truth also matters, as labels for training directly affect the models' learning of a particular phenomenon and the results of ML algorithms [135]. For instance, Kim et al. [77] found that third-party annotators had a significantly different perspective on bullying-related risks compared to the victims themselves. Annotators' background and expertise influence how the data is annotated for ground truth; thus, annotators should be selected carefully and their background should be stated clearly [5]. Essentially, such considerations interrogate whether the data being leveraged is ecologically valid for detecting the targeted risk for the desired user population. When there are multiple people labeling the same dataset, it is important to measure the Inter-rater Reliability (IRR) with at least a subset of the corpus as well as formulate ways to resolve any annotation disagreements between the annotators [87].

3.2 Considerations for Computational Models Using a Human-Centered Lens

Models, from a computational perspective [104], are the artifacts that encode decision or prediction logic trained from the training data, the learning program, and frameworks. The model learns from a set of features; a feature is a measurable property or characteristic of a phenomenon being observed to describe the instances. From a human-centered perspective, it is important to see if the computational models being built are human-informed and evidence-based in terms of drawing from risk models based on human theories [17]. Baumer et al. [17] introduced human-centered approaches to machine learning that leverage theoretical frameworks derived from and validated within the social sciences, such as psychology and communications. So, the main question to be asked for reviewing computational risk detection research revolves around "Are the algorithms grounded in human theory, understanding, and knowledge?". When reviewing computational risk detection models, it is important to look into explanations and the transparency of the model to understand if it could be accountable for making fair and unbiased decisions by utilizing human-centered approaches [96]. Reviewing explanations for different parts of the models includes, but is not limited to, identifying the types of algorithms or techniques that were used and why, how the algorithms were trained, what the input characteristics were, training parameters, fairness constraints and potential biases, features, and if the intended use of the output of the model and the choice of output structure was stated [59]. More importantly, depending on the task that the model is trying to do, especially in the area of risk detection, the outcomes of the model could be reviewed from different human angles; for instance, looking at the timing of the detection to

understand if the model is capable of detecting a risk before or while it happens, or if it needs full data to detect the risk after-the-fact [89].

3.3 Human-Centered Considerations for Model Evaluation

Once a machine learning model is developed, one must test the model to evaluate the performance and speculate the usage of the model. From a computational perspective [104], generalization error is defined as the expected difference ratio between the real conditions and the predicted conditions of any valid data. Usually, computational scientists test the models based on computational metrics such as accuracy, precision, or recall. Reporting precision and recall is important based on the application that the risk detection system will be used. Although such metrics provide a picture of the capabilities of the model, they are far from sufficient for evaluating the model. From a human-centered perspective, researchers could ask "How well do the algorithms perform computationally and in meeting end users' needs??" HCI design methods such as user studies could be used to evaluate the model from the users' perspective. User studies could provide valuable insights from stakeholders of the system including end-users. Yet, testing is important in the life cycle of deploying a machine learning system. Zhang et al. [165] summarized techniques for testing ML systems from a computational perspective such as testing properties (e.g., correctness, robustness, and fairness), testing components (e.g., the data, learning program, and framework), testing workflow (e.g., test generation and test evaluation). They mentioned before deploying the model online, conducting offline testing, such as cross-validation, to make sure that the model meets the required conditions is necessary. After deployment, predictions of new data can be analyzed via online testing to evaluate how the model interacts with user behaviors.

3.4 Human-Centered Applications of Computational Risk Detection Models

A system-based artifact is considered a primary research contribution within the SIGCHI community [159]. It is not until these computational models are tested in the wild, that we can truly determine their effectiveness. Thus, it is crucial that researchers not only develop and evaluate algorithms but also integrate them into applications so they can be evaluated by humans in real-world settings [149]. It is at this stage of development that the explainability of AI/ML systems comes to the light and how inclusive and interactive designs are used to create applications to make fair and inclusive decisions [166]. Especially in the context of risk detection, it is important to know how the model is integrated into a system, as there is an important ethical distinction between using risk detection algorithms for the purpose of surveillance and criminal justice [164] versus victim empowerment and advocacy [74]. After risks are detected, the next logical step is towards risk mitigation to reduce long-lasting harm [121].

In summary, these are different components of computational risk detection that we consider in our systematic review of the online sexual risk detection literature. However, we also believe that researchers could use this framework for conducting systemic reviews of computation risk detection literature more broadly. Next, we describe how we applied this human-centered lens when operationalizing the coded dimensions of our systematic review.

4 METHODS

Below, we describe how we scoped our literature search and systematically reviewed the articles included in our dataset.

4.1 Systematic Literature Search

For our initial literature search, we identified five electronic databases that included computational and interdisciplinary research on sexual risk detection (i.e., IEEE Xplore Digital Library, ACM Digital

Library, ScienceDirect, Springer-link, and ACL Anthology) to ensure comprehensive coverage of the relevant literature. We used combinations of the following keywords: sexual predator, sexual predation, sexual risk detection, sexual abuse detection, sexual grooming, sexual assault, online grooming, cyberpedophile, pedophile, paedophile, sex trafficking, predatory conversations. We included words like detection, recognition, and machine learning to find computational articles, as opposed to articles that studied the phenomenon itself from a more qualitative perspective (e.g., focus groups or interview studies). We explain our scoping criteria and relevancy coding next.

4.2 Scoping Criteria and Dataset Creation

Our initial search resulted in 296 unique papers. Next, we examined the paper title, abstract, keywords, results, and conclusions to identify relevant studies that met the following inclusion criteria

- The study was a peer-reviewed published work.
- The study was published between 2007 and 2020. We included papers that were published in these years, but there were not any papers that met our inclusion criteria before 2007.
- The study focused on sexual risk detection (our definition of sexual risks includes sexual predation, sexual grooming, sexual assault, sexual abuse, sex trafficking, and sexually abusive conversations).
- The study contained a computational/algorithmic approach or a system architecture on text and multi-modal data (including Natural Language Processing, Machine Learning, etc.)

We marked a paper as relevant if it met our relevancy criteria, which resulted in 57 relevant articles. Then, we cross-referenced the citations of these relevant papers to identify additional papers to include that may not have been included in our initial database search. This cross-reference exercise resulted in 116 unique papers (62 after removing duplicates from our initial search) in which 15 papers met our inclusion criteria. We did one more iteration of this search process, which identified two additional relevant papers. Having reached an apparent saturation point, we concluded our search with a final set of 73 articles for our review.

The primary reason papers were excluded (around 80%) was because the article did not specifically address sexual risks. These papers included cyberbullying detection [133], authorship attribution [70], and general privacy [14] and forensics papers [16]. The second most common reason for exclusion (around 15%) was that the study did not include a computational approach for sexual risk detection, such as papers that qualitatively assessed users' online sexual risk experiences [127]. We also excluded papers on image and video sexual risk detection papers because Tariq et al.'s [149] recent literature review on these approaches. However, we did include multi-modal approaches that included textual data. Next, we describe how we synthesized the literature.

4.3 Data Analysis Approach

We leveraged the human-centered lens proposed above (Figure 1) to code the 73 relevant papers. Our final codebook is presented in Table 1. We used an iterative process to refine our codes and allowed codes to overlap and be double-coded. Three coders labeled the same 15% of articles, and we calculated Fleiss's Kappa IRR [45], which is an extension of Cohen's kappa for three raters or more. This resulted in Fleiss's Kappa ranging from a substantial (0.70) to a complete agreement (1.00) [45] for all codes across dimensions. To resolve conflicts, the researchers discussed the articles until a consensus was reached and updated the operationalization of the codes for consistency. Then, the remaining articles were divided among the three coders. The first author reviewed the final codes to identify emerging themes, patterns, and potential gaps within the literature.

Table 1. Codebook

Categories	Dimensions	Codes (Subcodes)
Overall Characteristics	**Risk Type**: *What are the risk type and the key aspects of the risks that researchers are detecting?*	Sexual Grooming (75%), Sex Trafficking (12%), Sexual Abuse/Harassment (12%)
	Timing of Detection: *When is the timing of the risk detection?*	After (93%), During (38%), Before (5%)
	Target Person: *Who is the person that the system tries to detect? What is the age range of the target person?*	Perpetrator (80%) (Child, Adult, Posing as Child), Victim (17%) (Child, Adult, Posing as Child)
Data	**Data Source**: *What is the data source and data type? What is the privacy level of the dataset?*	**Dataset Source**: Perverted Justice (30%), PAN-2012 (22%), Combined Chat datasets (14%), Social Media (11%), Advertisements (11%), Queries (5%), Private Chat data (4%), SafeCity (4%), Games (3%), Forums (3%), Anonymous Platforms (3%), Blogs (3%), Generated by Participants (1%) **Data Types**: Text (64%), Meta Data (23%), Images/Video (11%) **Privacy Level**: Public (82%), Private (15%)
	Ground Truth: *How was the data annotated for training datasets? Did they report annotators' IRR for more than two annotators?*	**Annotators**: Existing Labels (44%), Outsiders (33%) (Researchers (26%), Moderators (3%), Clinical (1%), Crowdsource (1%)), Automatic Approach (23%), Insiders (5%) **Annotators IRR**: No (95%), Yes (5%)
Models	**Feature**: *What were the features used in the development of the model?*	Textual (85%), User (27%), Time/Location (19%), Semantic (16%), Style (15%), Behavioral (14%), Keyword Extraction (14%), Syntactic (11%), Sentiment (10%), Images (7%), Network (5%), Relationships (4%), Topic Modeling (4%)
	Algorithmic Approach: *What machine learning model(s) were used for the task? What is the detection task granularity for the studies that are on conversations?*	**Approach Types**: Machine Learning (83%) (Traditional (66%), Deep Learning (15%), Ensemble(2%)), Handcrafted/Rule-based (18%), System Architecture (7%), Graph/Network-based (7%) **ML Model Type**: Supervised (72%), Unsupervised (13%), Semi-Supervised (5%) **Algorithm Name**: Support Vector Machine (37%), Bayes (22%), Neural Networks (21%), Regressions (14%), etc. **Task Granularity**: Users (36%), Conversation Type (21%), Patterns (12%), Lines/Parts of Conversations (11%), Risk Levels (3%)
	Output Type: *What is the format of the model output?*	Binary (60%), Multi-class (33%), Clusters (10%), Predatory Stages (7%)
Evaluation and Application	**Evaluation**: *How was the model evaluated? Did they do user studies to evaluate the model by stakeholders?*	**Algorithmic Metrics**: Accuracy (53%), F1 (37%), Precision (23%), Recall (19%), etc. **User Study**: No (96%), Yes (4%)
	Artifact and Intervention: *What is the final artifact developed? Did the artifact provide any interventions of risk mitigation in addition to detecting risk?*	**Artifact**: Algorithm Only (86%), Chat System (1%), Forensic Investigation Tool (1%), Mobile App (1%) **Intervention**: No (93%), Yes (7%)

The definitions of our coded dimensions, local research questions, and grounded codes that emerged from our data are shown in Table 1. For instance, for Ground Truth, we coded for the annotation of the datasets and the humans' involved in the annotation process. Our codes include "Outsiders" (Someone other than the victim or the one that shared the story), "Insiders" (Victims or who shared the story), "Auto" (Used an automatic approach to label the data), or "Existing" (The paper used an existing labeled dataset). Task Granularity referred to the detection task granularity level for the studies that are on conversations, since there are several ways that researchers can implement risk detection on conversations (i.e., if the paper detects the "Level" of risk in a conversation, identifies the "Lines" of a conversation that are risky, identifies risky or predatory "Patterns" in conversations, or identifies Users as predators or victims in a conversation, or classifies the whole "Conversation" into different categories). Next, we present our results regarding our analysis of the categories according to the codebook and findings from the literature.

Table 2. Risk Types

Risk Types: Definition	Count (Percent)	References
Sexual Grooming: Detecting sexual grooming of an online predator. Grooming is a process to approach, persuade, and engage a child, the victim, in sexual activity by using Internet as a medium [110].	55 (75%)	[7, 9, 11, 22–24, 26, 27, 29, 30, 35–37, 40, 40, 42, 43, 46, 50, 51, 68–70, 76, 79, 84, 85, 90–92, 95, 97, 98, 101, 114–118, 122–124, 130, 131, 137, 140, 150, 151, 153, 163, 167]
Sex Trafficking: Is the process of recruitment, harboring, transportation, provision, or obtaining of a person for the purpose of a commercial sex act [3].	9 (12%)	[63–66, 81, 86, 141, 142, 157]
Sexual Harassment/Abuse: Includes online shared content of a range of sexually aggressive or harassing content, sexual assault, gender violence, sexual violence such as stories shared in the hashtag MeToo movement [56].	9 (12%)	[49, 54, 73, 75, 88, 136, 145, 147, 160]

5 RESULTS

In this section, we present our findings based on the 73 papers that we reviewed. We organize and present the results by our code dimensions in Table 1.

5.1 Overall Characteristics of Articles

Below, we report the descriptive characteristics of the articles in our dataset, including the risk types studied in the literature over time, timing of detection, and target (i.e., person) of risk detection.

5.1.1 Sexual Risk Detection Types Overtime. As illustrated in the Table 2, the research papers that we reviewed mainly considered four kinds of sexual risks: sexual grooming (N=55, 75%), sex trafficking (N=9, 12%), and sexual harassment and/or abuse (N=9, 12%). As shown in Figure 2, we observed some notable trends. Starting from 2007, with the rise of using social media and online chat rooms [112], researchers start working on combating the emerging problem of sexual grooming. In 2012, we saw a spike in the literature, likely due to the PAN12 competition for predator detection [67]. Concurrently, sex trafficking detection literature emerged in 2012, which coincided with the announcement of the Obama administration efforts to combat human trafficking [1]. Meanwhile, we saw the rise of the #MeToo movement at the end of 2017 with social media users sharing sexual harassment self-disclosures that went viral [58]. Consequently, this coincides with the computational research on sexual harassment and/or abuse that emerged in 2018. Further, the significant increase in sexual risk detection publications in 2019 highlights the critical importance of computational sexual risk detection and the need for a systematic review that synthesizes this body of research in a way that can create a cohesive research agenda moving forward. There seems to be a slight decrease of literature in 2020, this might have been caused by several underlying factors. We will further discuss these factors in our discussion section.

This trend in the past literature represents the algorithmic advances on sexual grooming detection which is an important issue to tackle. But also demonstrates the overlooking of critical sexual risk types that should carry the same weight, if not more, in some cases, such as unwanted sexting and sexual abuse. The long-lasting impacts on victims, as well as the distinctive characteristics of relationships that engender sexual risk (often someone close to the victim like a family member) [127] together underscore a necessity to close this gap in examining sexual abuse online.

5.1.2 When Risk Detection Occurs. Most algorithms in our review focused on after-the-fact risk detection (N=68, 93%), which examines ways to detect the risk with an underlying precondition after online risky behavior has occurred. For instance, Ringenberg et al.'s [130] machine learning model to differentiate contact-driven and fantasy-driven sexual solicitors were based on the dataset of complete conversations. Yet, there was a smaller subset of studies (N=28, 38%) that were capable of detecting risks during the occurrence through detecting predatory patterns and risk levels

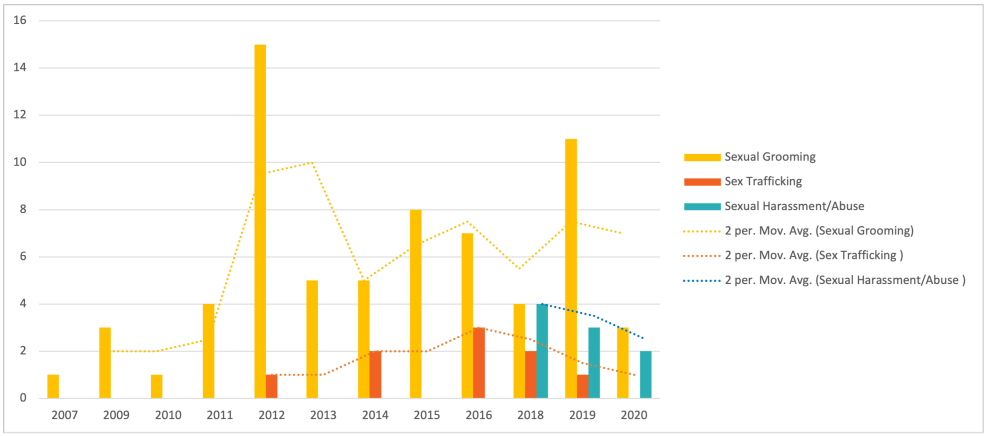

Fig. 2. Number of Publications by Risk Type Over Time

within a conversation [23, 24, 26, 27, 30, 35, 37, 38, 42, 43, 50, 51, 63, 69, 79, 84–86, 95, 101, 115, 123, 131, 136, 140, 151, 163, 168]. For instance, Cano et al. [30] sought to detect three online grooming stages: 1) Trust Development; 2) Grooming; 3) Seek for physical approach, while sexual grooming is happening. Only 4 (5%) papers detected sexual risks before and during the event [81, 92, 99, 118]. MacFarlane et al. [92] detected personal information in children's online chats and proactively blocked private information from being sent to prevent victimization. Kostakos et al. [81] built a model to predict risk factors of users that depict the likelihood of being drawn to online sex work and illustrated a potential methodology that could be used to identify ones with high-risk factors. The way that the data is given to a computational model as an input affects the fact on when the prediction result of the model. In summary, the automated models for sexual risk detection of the past literature have generally been based on after the risk occurs. While detecting online risky interactions after-the-fact could help law enforcement agencies detect sexual predators, at this point it may be too late to prevent victimization.

5.1.3 Whom is the Object of Sexual Risk Detection. The majority of the papers focused on identifying predators (N=59, 80%) or conversations that included predators. Another 13 (17%) papers [29, 49, 54, 63, 73, 75, 81, 88, 92, 142, 145, 147, 160] focused on identifying victims. The focus of these papers was mostly to identify social media self-disclosures about sexual harassment and abuse (N=11, 15%). For example, Hassan et al. [54] leveraged the hashtag #MeToo movement, where women shared their stories of sexual violence in social networks, to collect and to classify the sexual harassment reports among the posts.

5.1.4 Age Range for Sexual Risks Detection. Most computational approaches to sexual risk detection were based on predators who were adults and victims who were adult volunteers posing as children (predator/pseudo-victim) (N=40, 55%) given the prevalent use of the Perverted Justice and PAN-12 datasets. One paper [168] created a chatbot posing as teens to talk to pedophiles. There were 20 (27%) papers that either did not specify the age of predators or victims or used data from adults. As an instance of not specifying age, Karlekar et al. [73] focused on identifying sexual harassment posts on the SafeCity dataset without differentiating users by age. Although some posts were about the child abuse disclosures, they do not directly identify child victims. In fact, in some cases, an adult shared a story about a child. These papers were indifferent to main dissimilarities between data from underage users and adult users. Only a few (N=9, 12%) papers [63, 69, 85, 92, 99, 115, 142, 145, 157]

focused on underage victims, which were mostly proposing chat systems or crime investigation tools, but not proposing models on chat conversations. A unique study in terms of doing an effort to focus on youth's data is the study by Roy et al. [136] which used hypothetical situations and created a dataset by recruiting youth and gave them scenarios to create abusive text messages given each scenario.

5.2 Assessing the Ecological Validity of the Data (RQ1)

In this section, we present our findings regarding the data and ground truth annotated training datasets used for sexual risk detection.

5.2.1 Data Type: Primarily Focused on Text. The majority of the papers (N=47, 64%) relied on text data for their risk detection models, (media types for each reference is displayed in Table 3 in the Appendix). Although the datasets for these papers include metadata, for example, PJ has timestamps of the messages, these papers did not utilize this additional data. There were some papers (N=16, 21%) that incorporated metadata in addition to text such as time or user profile data. For example, Potha et al. [123] used time series modeling to reveal conserved temporal patterns or variations in the strategies of predators in the PJ dataset. As such, approaches that focused on text often did so in the absence of images (N=66, 90%) or other types of multi-modal data. A few papers (N=7, 10%) considered images, text, and meta-data. Most of these papers (86%) were trying to detect sex trafficking from online advertisements, and there were no papers using images and multi-modal data for sexual risk detection within conversations. We found 2 (3%) papers that solely used metadata, such as profile features. Only one paper [151] utilized both text and images to detect adult content on Facebook posts. Overall, as we included any articles that included text plus other data types, we identified a gap in the literature for analyzing multi-modal and metadata such as user and temporal features in order to acquire the context of the online interactions to build effective detection models for sexual risks.

5.2.2 Data Sources: Mostly Publicly Available and Public-facing Datasets. We analyzed papers in terms of their datasets' source, type, and privacy. Our analysis uncovered that papers mostly worked on public datasets (N=60, 82%), and only a few (N=11, 15%) were using private datasets [7, 36, 37, 69, 75, 98, 118, 136, 145, 147, 150]. Table 4 in the appendix illustrates the datasets of the papers examined in this literature review. Private datasets are mostly scraped from Social Media (e.g. Twitter), but some kept the platform name anonymous and some were private chat data (e.g. Whisper [150]). Some papers used several datasets for their analysis, for instance, Pandey et al. [115] used social media data, blog posts data, and data from forums and public chats.

Next, we discuss the dataset type, name, and source based on the prevalence of datasets. We identified that based on risk types that these papers are trying to detect, the sources of the datasets have commonalities. The most popular public datasets used in the literature for identifying sexual groomers include Perverted Justice (PJ) dataset [47] (N=22, 30%) and dataset from PAN-2012 competition (N=16, 22%) [67]. A few papers (N=10, 14%) used combined chat datasets with PJ chat data; for instance, Rangel et al. [125] used the PJ dataset and the dataset for author profiling at PAN 2013. Pranoto et al. [124] used scripts from www.literotika.com which contains sexual conversations shared by adults in a legal manner. Overall PJ and PAN-2012 were the only public datasets that helped researchers advance computational solutions for identifying sexual predators.

Moreover, some studies (N=8, 11%) used data from social media, for instance, Suvarna et al. [147] sought to identify sexual assault victim-blaming language on Twitter scraped posts. They justified their decision of creating a dataset using Twitter by explained that other platforms such as Reddit's or Facebook's community affordances may not provide a structure for people to voice their actual opinions for sensitive topics such as blaming the victim due to the presence of moderators. Datasets

from online advertisements were all used by sex trafficking papers (N=8, 11%) except one paper on sex trafficking [81]. This paper, utilized data from a popular European adult dating forum to collect data for each user's profile page and qualitative data on users' self-reported behavior regarding paid sex. In addition to the dominant trend of a few public datasets and social media data, there were noteworthy papers that used different datasets; queries in P2P systems or networks (N=4, 5%), private chat datasets (N = 3, 4%), and a public dataset named SafeCity (N=3, 4%) were seen in papers. All sexual harassment and abuse detection studies are mostly based on data scraped from social media and the SafeCity dataset, with the exception of a study by Roy et al. [136]. For instance, Khatua [75] extracted 0.7 million tweets with the hashtag "MeToo" for the social movement against sexual harassment. Due to the lack of publicly available datasets that are representative of users and the difficulties to collect real digital trace data because of the sensitive nature of this problem space, Roy et al. [136] recruited participants to generate abusive text messages that could be used for classification. The participants were given abuse scenarios and asked to create corresponding text messages. One of the limitations related to this dataset was that there is not enough data to capture all the features related to dating abuse. Overall, we observed some good practices among the literature for justifying their choice of dataset, but the datasets were heavily skewed towards the PJ dataset that is not based on real-world users' conversations. Therefore, these conversations cannot be considered ecologically valid for the basis of risk detection. In addition, most of the time risks occur in private conversations [161], the primary use of publicly scraped datasets limits their usefulness in addressing the actual problem.

5.2.3 Ground Truth Annotations: Reliance on Third-party Annotators. Our finding on how papers provided ground truth labeling for ML training datasets and who annotated the data unpacks that most of the studies relied on existing annotations that were done mostly by researchers without relevant domain expertise, this is illustrated in Table 5 in the appendix. We found that a noteworthy proportion of the papers that we analyzed were based on existing labeled datasets (N=32, 44%). These datasets were mostly either PJ or the PAN12 dataset (N=23, 32%). Of the research papers that we reviewed, there are 45 papers that labeled their own dataset, most of which relied heavily on third-party (N=24, 33%) rather than the person who experienced the sexual risk. We further examined the papers that relied on third-party annotators for the data, particularly focusing on whether the papers explicitly included descriptive data about the annotators such as their backgrounds, expertise, and if any types of training was provided to them. Most third-party annotators were researchers (N=19, 26%) that did not appear to have special expertise on the matter, and only one paper had annotators that were clinically informed or experts. Chowdhury et al. [49] expressed that they had three independent annotators from Clinical Psychologists and Academia of Gender Studies for annotating their entire dataset of self-disclosure of sexual harassment tweets. The authors also provided their annotation guidelines about the types of posts that were considered sexual harassment and the annotation process.

There were only a few papers (N=4, 5%) that took into account the perspective of the individuals who experienced the sexual risks. These papers share a common characteristic of being on social media posts or tweets of self-disclosures of sexual harassment [49, 75, 145]. The other paper by Kostakos et al. [81] selected a European adult dating forum for data collection and used social data mining to collect quantitative data for each user's profile page and covert online ethnography to collect qualitative data (interviews) on users' self-reported behavior regarding paid sex. They labeled user profiles based on the interviews conducted with users about their tendencies to buy and/or sell sexual services in the forum. This dataset was used to predict the risk factors of a larger poll of users. We found that only 4 (5%) papers [49, 79, 88, 136] reported Inter-rater Reliability (IRR) when they had more than one annotator to measure the quality of agreement between annotators.

Liu et al. [88] and Chowdhury [49] articulated the annotation procedure and used Cohen's kappa coefficient of inter-rater agreement for labeling sexual harassment stories, while Roy et al. [136] stated using Light's Kappa; an inter-rater agreement statistical consistency measure. Kontostathis et al. [80] reported their IRR using Holsti's method for annotating Perverted Justice's conversations.

In addition, some papers used automated methods such as keyword match or classification software for the annotation of their datasets (N=17, 23%). For instance, Michalopoulos et al. [98] used software for document classification that performs instant classification of a given text. Kontostathis, et al. [79] built a keyword-based software named ChatCoder to label and analyze predation chats and to identify a luring category for each. To make it more suitable to the online context, they adopted a simplified version of the Luring Communication Theory (LCT), a framework proposed by Olson et al. [110] that describes the process child sexual predators use to lure their victims into a sexual relationship. The study provided extensive details on the labeling process such as how the codebook for the labels was developed as well as the coding process stages. In summary, we found the reliance on existing data labels among the literature in this area and the ones who labeled the data were mostly by researchers.

5.3 Models Grounded in Human-Centered Theories and Knowledge (RQ2)

In this section, we provide our results of the review about models and their characteristics including features used by models, approach type and model selection, and model output.

5.3.1 Feature Selection: Textual versus Behavioral Features. The most commonly used features by papers were textual or lexical features (N=62, 85%), which could be drawn from raw text or statistics of the input text, however, theory or behavioral driven features are rare among the literature (illustrated in Table 6 in appendix). The textual features include n-gram use, character modeling, bag-of-words (BoW) models, term-frequency-inverse document frequency (TF-IDF), word embeddings, skip grams, Hashing Vectorizer, the length of the text, count/ratio of "emoticons", count/ratio of profanity, and the number of pronouns. For instance, some approaches used the number of reframing verbs (e.g., teach, practice) or the number of desensitizing verbs (e.g., touch, kiss) [95]. We also noted deep learning approaches that aimed to model the language in the text. Neural networks such as works by [35, 42, 76] used a bag-of-words representation that summarizes a conversation as the number of occurrences of each word in the vocabulary, regardless of the order in which they appear and do not require an explicit set of features.

User-based features, which are characteristics of a user's profile that can be used to make a judgement on the role played by the user in an online exchange and include age, gender, and sexual orientation, etc., were also common (N=20, 27%). Some papers considered time or/and location (N=14, 19%) as features. For example, Elzinga et al. [41] analyzed the change in the relationship between the offender and the victim over time to further examine how the threat level changed over the course of the conversation. Semantic features (N=12, 16%) represent the basic conceptual components of meaning for any lexical item. For instance WordNet [44] classifies words and uses hyponymy and hypernymy to establish semantic relationships between synsets that some papers such as the work by Iqbal et al. [69] used. In some papers (N=11, 15%), we saw the use of linguistic style or metrics of linguistic complexity that can help the model to distinguish language from the writing style of users. These metrics include Linguistic Inquiry and Word Count (LIWC) (vocabulary lists known to help measure emotion in text) [119], readability, coherence, perplexity measures, and subjectivity measures. The keyword extraction method was used by 10 (14%) papers to extract keywords, it included named entity recognition, which involves extracting entities (names, location, email addresses), word clouds, or tag clouds are another example of keyword extraction. Next, there were some other features that were used by the literature less frequently including, syntactic

features (N=8, 11%), sentiment features (N=7, 10%), image features (N=5, 7%), network features (N=4, 5%), topic modeling (N=3, 4%), and relationships (N=3, 4%).

Behavioral features of predators or victims were also utilized (N=10, 14%); Vartapetiance et al. [153] used the Cycle of Entrapment (deceptive trust development, sexual grooming, isolation) from the Luring Communication Theory (LCT) [110] to differentiate predators by the process of entrapment used by child sexual predators to lure their victims into sexual relationships. Perpetrators usually approach the victim to build not only sexual but also emotional relationships [110]. Similarly, another study included as feature, fixated discourse, which is the unwillingness of the predator to step out of the sex-related conversation even if the potential victim wants to change the topic, as a feature [22]. These last set of features are a good example on how to leverage the knowledge from other fields that have studied the different stages and processes behind online sexual grooming to guide the feature engineering process. We observed that papers used a variety of features for detecting sexual risks and among them textual features were most popular.

5.3.2 Algorithmic Approaches: Mostly Traditional Supervised Machine Learning. We observed trends in terms of approaches used by sexual detection literature (demonstrated in Figure 3). Traditional or classic ML approaches were used in earlier studies in the area of pedophile detection by researchers. After Traditional ML approaches, system architectures and hand-crafted/rule-based models came to the picture. In the latest years, deep learning models have been used in the literature more recently since 2016, and in the last year an ensemble approach was proposed.

The majority of the studies relied on traditional or classic ML algorithms such as Support Vector Machines (SVM) (N=48, 66%) (Approach types by references are displayed in Table 7 and algorithms' names are in Table 8 in the appendix). For instance, Pendar et al.'s study [117] was one of the first works that used SVM and K-Nearest Neighbors (KNN) on the PJ dataset. Fauzi et al. [43] found that soft voting based ensemble for distinguishing predatory conversations from the normal ones performs the best on the PAN-12 dataset compared to separately using Naive Bayes, SVM, Neural Network, Logistic Regression, Random Forest, KNN, and Decision Tree. That said, the study found that Naive Bayes outperformed other classifier models for differentiating between a predator and a victim in predatory conversations. Despite the increasing body of research that incorporates deep learning models, we were only able to identify a relatively small subset of the corpus that used deep learning in their approach (N=11, 15%). Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Long Short-term Memory (LSTM) methods have been used. Most of these models were for detecting sexual harassment, abuse, or sex trafficking (N=6, 8%) rather than sexual grooming detection (N=3, 4%). In one of the studies, Misra et al. [101] encoded predatory behavior purely from style (characters) to do authorship attribution on PJ corpus using a CNN model. One of the reasons that deep learning models were used less in the literature might be due to the lack of large datasets since most deep learning models are known to require large amounts of training data. For instance, for sexual harassment detection using deep learning, researchers used publicly-available dataset SafeCity which includes 9,892 stories [73, 88, 160].

The papers that used ML or deep learning models mostly used supervised algorithms (N=48, 66%), 6 (8%) using unsupervised, 4 (5%) using semi-supervised, and 4 (5%) using both supervised and unsupervised. Given the detection or prediction problem specifications, it is normal to find the supervised approach as the dominant selected type since in supervised approach, an input including the data labels is required for the models to be able to detect or predict the object of interest (users or types of conversations). We also noticed that quite a number of papers used semi-supervised approaches that only required a small amount of labeled data along with a huge dataset with a few number of labels as input for the models. Although a Semi-supervised approach can help address either the lack of available annotated datasets or the lack of annotators for a huge

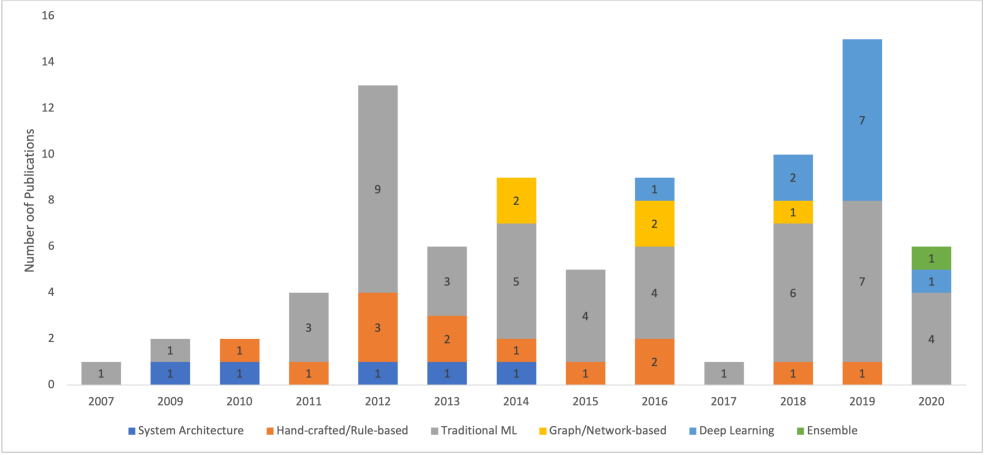

Fig. 3. Frequency Distribution of Approaches in Reviewed Papers Over Time

dataset, sometimes the models do not detect or predict accurately. For example, Kostakos et al. [81] used a semi-supervised learning approach with data collected from a popular European adult forum. Only 78 users were labeled based on their risk evaluation and left out 28,832 users for the model to predict their risk assessment, which their best model yielded only 79%. Unlike the two previous approaches, unsupervised approaches do not require annotated data as the models' input. With an unsupervised approach, models can capture patterns and extract knowledge from the input through algorithms, in most cases, clustering algorithms. For example, Toriumi et al. [150] was capable of uncovering risky communication behaviors on private chat systems used by minors to provide effective monitoring based on active communications.

Other than ML approaches hand-crafted or rule-based models were also commonly chosen as the methodology (N=13, 18%). For instance, Vartapetiance et al. [153] approach was built heuristically based on the Cycle of Entrapment which we discussed earlier. They created several categories for classifying conversations based on keywords presented and defined thresholds for assigning conversations to those classes. A few studies (N=5, 7%) used graph and network algorithms, in which all were detecting sex trafficking. Ibanez et al. [66] demonstrated that the content available in online escort advertisements can be used to identify provider networks and potentially roles using network graphs. We noted a lack of graph and network analysis for other types of risks such as sexual grooming or harassment on social media where considerable risks happen by analyzing social networks of users. A few papers proposed system architecture or tools (N=5, 7%) that were mostly based on filtering and blocking certain information that might cause risks for minors. Some (N=3, 4%) of these papers focused on sexual grooming [85, 92, 118]; for instance, [92] proposed architecture for detecting intent, time and location to block messages when children chat online. The other two papers focused on sex trafficking [142, 157]; for example, Silva et al. [142] proposed a system that filters and retrieves textual and image data related to sex trafficking to help law enforcement agents. On average, these systems were proposed in the year 2012 and they only block certain information such as address or phone numbers.

5.3.3 Task Granularity: Primarily Conversation Level Detection of Predators. Our results about the granularity level of the risk detection tasks for models on conversations (illustrated in the Table 9) reveals that most of the studies that we reviewed focused on detecting and differentiating predator users and victims (N=27, 36%), rather than identifying patterns and changes that are

indicators of risks. Some studies tried to identify predatory conversations (N=15, 20%). Kim et al. that [76] attempted to first classify each message and then used those results to classify the entire conversations using RNN to overcome the high potentiality of having large blocks of conversations which might include hundreds of messages. Overall researchers were successful in the task of differentiating users with good algorithmic performance metrics. Some studies aimed to identify patterns (N=9, 12%) or characteristics/structures that contribute to being a predator in conversations. These patterns were usually associated with how predators approach victims and common methods used by them driven from theoretical frameworks for sexual grooming. Zambrano et al. [163] used existing theoretical frameworks to map different stages of the life cycle of the grooming (including gathering information, gaining access, lateral movement, escalating privileges, execution, debrief) within conversations of predators. They framed grooming as a vector of attacks which could be used for determining patterns of malicious behavior online. A few reviewed papers tried to identify predatory lines (N=8, 11%) in the conversations. Bours et al. [27] looked at individual messages to determine if such a message belongs to a sexual predator or not. While only two studies (3%) [131, 140] tried to identify the predatory risk levels. Ringenberg et al. [131] used Fuzzy Sets for labeling messages for three levels of risks (low, medium, high), and developed a NN model that uses these fuzzy membership functions of each line in a chat as input and predicts the risk of interaction.

5.3.4 Algorithmic Output: Mostly Binary. Our analysis of the output types of algorithms is summarized in the Table 10 in the Appendix, we list the publications grouped by the specific classification setup used. In there we show that most papers used a binary classification setting (N=44, 60%) where they usually classify conversations as predatory or not predatory, and users to predator or victims. Several studies performed multi-class classification (N=24, 33%) by differentiating class type rather than trying to fit all different types into binary classification. Khataua et al. [75] classified different types of sexual violence and associated risk of them to 4 different categories based on the locations that it occurs. Some papers (N=7, 10%) tried clustering methods, for instance, Kontostathis et al. [79] aimed to cluster different types of predators via their language pattern usage using K-means on the PJ dataset, and found 4 clusters produce the best results, meaning there exist four different types of predators. A few studies identified predatory stages (N=5, 7%), such as Potha et al. [123] where they examined the question set of each predator with a view to capturing the tone of predator's defensive or aggressive questions in order to identify patterns of predator's behavior that can be generalized in a real-life conversation as time series, i.e. using windows that only capture a short period of the predator's attacking strategy.

Overall, in this section, we reviewed the models and the specification for these models. In the next section, we will report our findings about the artifacts that were built by the reviewed studies and will discuss how they evaluated the models.

5.4 Evaluation and Performance Metrics (RQ3)

In this section, we discuss our results regarding the evaluations of the performance of the artifacts produced from the reviewed studies.

5.4.1 Evaluation: Focus on Numerical ML Performance Metrics. For the evaluation of the performance of the models, most articles that we reviewed focused on computational performance (N=57, 87%) and did not leverage user studies or human evaluations to assess their approach. Only 3 papers (4%) [54, 85, 118] performed user studies to evaluate their models. Latapy et al. [85] evaluated their tool for automatic detection of paedophile queries by human experts that work in law-enforcement agencies and well-established NGOs. The reviewed articles mostly reported for Accuracy (N=39, 53%), F measure (N=27, 37%), Precision (N=17, 23%), Recall (N=14, 19%), and other

numerical performance measures. Some papers only measured accuracy [51, 95], which is not enough for assessing the performance of a classifier without other measures as it might be biased toward the majority class [109], especially in this application in which the dataset is imbalance toward non-risky instances. PAN-12 was the first to benchmark performances of the models by the standard Information Retrieval measure of Precision (P), Recall (R), and F measure (weighted harmonic mean between Precision and Recall) [67]. Inches et al. [67] pointed out that the standard F measure equally weighted with precision and recall is not always desired. For the problem of identifying predators, detecting the right suspected users as predators is more important (precision) than having more suspects (recall). Since police officers would rather act fast towards the "right" suspect rather than "all" the possible suspects. Therefore, they used a measure of F with a factor equal to 0.5, to highlight precision. On the other hand, for the problem of identifying predatory lines in conversations, it is more important to retrieve more relevant lines (recall) to be used as strong evidence toward a suspect. So they used a measure of F with the factor equal to 3 to highlight recall. Overall, for the PAN-2012 competition, participants were able to detect the predators with accuracy of 93%. But for the second task of identifying predatory lines, researchers were not successful (47% accuracy) [67]. Researchers continued to improve these computational performances for these tasks even after PAN-2012. For instance, Kim et al.'s [76] recent study compared their results to the 16 competitors at the PAN2012 cyberpredator detection competition [67] and claimed that their results placed them first with respect to recall and F1 score, third with respect to F0.5 score, and fifth with respect to precision. However, they claimed that recall was the most important measure for the problem of detecting predators because it helps determine the fraction of predators who would go undetected, which is in contradiction of Inches et al.'s statement [67] that precision is the most important factor.

5.5 Application and System Artifacts (RQ4)

In this section, we discuss the artifacts and applications produced from the reviewed papers and the usability of these algorithms and artifacts.

5.5.1 System Artifacts: Mostly Algorithms Only. An emerging theme from the literature in this area is their focus on the development and improvement of risk detection algorithms which resulted in improving the computational results of algorithms (N=63, 86%) rather than creating a system-based artifact. A few studies (N=6, 8%) [41, 79, 118, 140, 142, 157] created Forensic Investigation Tools to facilitate analysis for law enforcement to detect predators in chat conversations or find predatory relationships. Laorden et al. [84] created a conversational agent (Negobot) that posed as a child in chats to detect conversations with paedophiles. The aim of this research was to gather enough evidence from users suspected as paedophile to help the authorities to detect and identify paedophile behaviours. Michalopoulos et al. [99] created an Android application that detects sexual exploitation attacks by capturing incoming SMS messages, processing it with the assistance of classification and clustering techniques in a distributed topology. Additionally, there is a lack of APIs that could be integrated into social media, chat rooms, forums, and other platforms for detection of online sexual risks.

5.5.2 Intervention: Focused on Detection without Mitigation. Most of the literature did not suggest or implement intervention strategies (N=68, 93%). Only 5 (7%) papers [92, 99, 118, 157, 168] provided some form of intervention strategies. For instance, as previously mentioned Michalopoulos et al. [99] developed a sexual risk detection Android application, which provided an intervention method to send a warning signal to the designated parent who is responsible for further actions in case of high risk. MacFarlance et al. [92] developed a prototype of an agent-mediated autonomous system to automatically detect and block the transmission of personal data when children chat

online, to detect and prevent attempts by users to arrange meetings. Most of the previous literature proposed systems for sexual risk detection without providing interventions for keeping users safe from online sexual risks.

6 DISCUSSION

In this section, we provide a summary of key trends we found within the literature, discuss the implications of our findings, and reflect on potential gaps and opportunities that we found in our review of the literature. We also provide directions for future research in the area of sexual risk detection. Then, we reiterate the need for human-centeredness in computational risk detection.

6.1 Trends, Gaps, and Opportunities for Future Research on Sexual Risk Detection

6.1.1 Sexual Risk Detection is Skewed toward Sexual Grooming. We found a larger proportion of studies focused on sexual grooming, rather than other sexual risk types. The skewed direction of studies on sexual grooming detection might be due to the PAN 2012 conference competition, which led to a significant increase in publications in those years. In addition, there might be other underlying reasons for the decrease of literature in recent years, such as the iterative process that we used for finding relevant papers by cross references citations and the emergence of a novel coronavirus that was first reported in late December of 2019 [55]. Therefore, one suggestion to bolster more research towards tackling other types of sexual risk detection would be to hold hackathons or similar competitions to have groups of researchers converge on these topics. In more recent years, we saw an emergence of work on sex trafficking, sexual abuse, and harassment, which we encourage the uptick in this trend. The long-lasting impacts on victims, as well as the distinctive characteristics of relationships that engender sexual risk (often someone close to the victim like a family member) [127] together underscore a necessity to close this gap in examining sexual abuse online. However, it would also be valuable to study factors that are pre-cursors to sexual violence, such as unwanted sexting requests. Sexting involves sending, receiving, or forwarding sexually explicit messages, images, or media to others through electronic means and is prevalent among adults [152] and is becoming prevalent among youth [106, 127]. Sexting involves inherent risks given the possibility of negative outcomes involving bullying, non-consensual dissemination of sexual imagery, or increased violence against women [82]. Therefore, sexual risk detection for sexting and other types of unwanted (or wanted, in the case of minors) could be an important future direction for the sexual risk detection literature.

6.1.2 Datasets that Reflect Real-World Interactions and Users are Needed. It is crucial to train models utilizing real-world data that are representative of the target users, in order for the approach to be usable in the wild [19]. Yet, we found that most studies were based on public datasets. For instance, datasets analyzed for sexual harassment or abuse were based on public posts on social media, such as Twitter. Yet, we assert that analyzing public discourse about harassment and abuse is not enough for tackling this problem, as most sexual risks occur in private channels [161]. Thus, future works should leverage conversations in private channels for training their models. We understand that the data collection efforts associated with this recommendation are not trivial. It would involve researchers to engage in public scholarship and recruitment efforts to reach sexual harassment and abuse survivors, convincing them to share the details (i.e., digital trace data) of their most intimate and traumatic sexual experiences with researchers who they do not know or trust. As such, we encourage researchers to be thoughtful in how they engage in scholarship with vulnerable populations, in a way that respects their dignity and their privacy [94, 156]. Moreover, collecting private data by researchers and privacy constrain of not sharing the data publicly pose more challenges in terms of replication of research for the research community. However, a way

to mitigate the concerns of replicability is to form coalitions of researchers who work together to solve important problems as a community of scholars who are committed to the protection of human subjects and data privacy. Another fruitful path of research would be to explore data anonymization approaches for ensuring the confidentially of participants when sharing datasets semi-publicly with other researchers through well-thought-out data sharing agreements and prior consent of research participants.

Further, the datasets reviewed in our paper were heavily skewed toward Perverted Justice (PJ). As the data acquisition for underage victims or law enforcement officers posing as children is challenging due to the laws and procedure involved [67], the PJ dataset focused on data from predators and adult volunteers posing as children. A limitation of this dataset is that it is not reflective of real children's interactions with predators, since the participants were adult volunteers. Approaches with datasets consisting of volunteers posing as children are potentially problematic, as it is possible to identify adults pretending to be children based on writing styles and authorship attributes [11]. While this approach may be adequate when the sole purpose is to detect predatory behavior, such an approach is not effective when the goal shifts to detecting behavioral patterns of youth victims. For instance, qualitative researchers found that adolescents struggle to handle sexual solicitation from people they know, but it is easier for them to reject sexual solicitations from strangers [127]. Therefore, taking a complementary approach of detecting strangers combined with detecting potential victim-signaling behaviors of teens could potentially bolster efforts to intervene and advocate for youth before victimization occurs. To do this, however, future research needs to create datasets that are ecologically valid for youth and use conversations from real adolescents to train such models.

6.1.3 Establishing Ecologically Valid and Trauma-Informed Ground Truth. Risk is a highly subjective construct, and defining and quantifiably operationalizing risk has substantial impacts on risk management and safety [120]. Through our literature review, we found that what denotes and marks sexual risk behaviors online has not been grounded in a systematic way. In most cases, the literature in our review did not report IRR to establish consistency of their ground truth annotations. Although theories were sometimes used for feature selection, such grounding was not used for data annotations. Theoretical design [17] utilizes a wealth of concepts and theories from behavioral and social sciences toward creating high-quality ground truth annotated datasets. For instance, future studies could leverage existing theoretical frameworks related to online sexual risks for establishing ground truth for data annotation. Computational scientists should form more collaboration with social scientists to build more frameworks for data ground truth for emerging sexual risk issues. The subjectivity of risk, in these cases, negatively impacts establishing robust ground truth [83] for sexual risk detection datasets used for ML risk classification. Researchers need to fill this knowledge gap by establishing an understanding *what* the most salient online sexual risk factors are through social science and psychological theories and by engaging directly with clinical experts. Given the highly subjective nature of risk, it is important that studies provide clear explanations on how they defined sexual risks for ground truth annotations, annotation procedures, and the annotators' backgrounds. It is essential to clearly explain the data annotation process, include the definitions that were used for data labels, consider annotators' demographics and expertise, and ensure sufficient inter-rater reliability, to address any potential bias from human annotators. These labels should be informative, discriminating, and independent. In data labeling, domain knowledge and contextual understanding help annotators to create high-quality datasets.

Including clinicians and subject matter experts in the annotation process would help improve the overall quality of ground truth. People have different perspectives on sexual risks based on their personal experiences and backgrounds; thus, it is important to look into annotators'

backgrounds and experiences. Subject matter experts, such as clinical psychologists or those who have background and training in sexual trauma, would be more equipped to annotate these datasets [5] than research faculty and/or their students. In addition, people who have personally experienced some form of sexual risk or abuse may have a better understanding of these types of risks than people who have not, making sexual abuse survivors another group of people suitable for the annotations. Yet, we make this recommendation with caution; asking past sexual victims to annotate sexual risks could inflict trauma by triggering memories of the annotator's own abuse experiences [143]. Therefore annotation process should be completed with utmost care and different techniques such as stress management, time management, relaxation, leisure, and personal renewal [162] with frequent mental health checks should be used to make sure the annotators with past trauma are not experiencing any difficulties and are willing to continue the annotation. Since the current literature relies heavily on third-party annotations, they might not have the same perspectives as people who were actual victims of online risks [77]; future studies should also take into account the perspectives of the victims. There are ways of taking into account the perspective of victims without having them directly annotate their data, for instance, Kim et. al's work [77] where they used user-labeled posts on a mental health peer support to indicate ground truth for bullying experiences. Yet, we acknowledge that resource constraints and time limitations are reasons why researchers relied heavily on non-expert annotators. One possible middle-ground to overcome these limitations is to have experts create the codebook and guide non-experts performing the annotations.

6.1.4 Models Need to Consider Contextual Information. We found that most of the literature disregards the context of the risky interaction, which can provide important information for defining risks. For instance, most approaches focused on a single data type such as text, in the absence of images or available meta-data. Due to challenges of multi-modal ML, such as representation, translation from one modality to another, alignment, fusion, and co-learning, most of the literature only focused on one modality [13]. Multi-modal data provides more context, especially in social media and online conversations. AI methods need to be able to interpret multi-modal signals together to make improvements in understanding the complexity of real-world experiences. Failing to include contextual data in the training of detection models is likely to result in a significant amount of false positives. A high false positive rate may cause systems to over-flag content, consequently reducing the system's capacity for effective risk mitigation. Although there are challenges in multi-modal approaches, improving the ability of models to process additional context can be a fruitful undertaking.

As many societal and psychological factors are at play when sexual risks happen, exploring mechanisms to model these factors in automated approaches becomes important, and we realize that there have been limited efforts in this direction. Although we observed good practices from the sexual grooming detection literature [22, 110, 153] on how to leverage knowledge and theories from other fields that have studied the different stages and processes behind online sexual grooming to guide the feature engineering process, these features have not been used as widely. Therefore, future works need to consider using these behavioral and contextual features. In fact, it would be interesting to explore if more recent approaches on representation learning are able to learn this type of latent information from the data. Further, some literature did not provide justifications for the features used. Although for some applications, the initial solutions for feature selection are often intuitive and based on human observation, there is still a need to perform posthoc analyses of features to show their usefulness as well as how they connect with established theories. For grounding future computational models in human theories and knowledge, more data needs to be empirically analyzed by experts from humanities and social science fields to create new theories

or frameworks that can be leveraged in AI models. For example, by analyzing social media data that involves risk, social scientists could create frameworks related to the ways in which youth may unintendedly indicate their vulnerability to become more susceptible to such risks. Thus, these frameworks could be integrated as features, inputs, or design of algorithms and models for detecting such instances. Since in some cases the risks phenomenon that happen online is so new that theories might not yet exist to inform technical development.

6.1.5 Advancing the State-of-the-Art in ML through Deep Learning. We found that most approaches are based on traditional ML algorithms rather than deep learning algorithms; this might be partially due to the lack of data, since deep learning models require large amounts of training data. Moreover, massive computing resources, and significant amounts of time are needed to successfully train deep learning models [33], and both of these items are often only available to big tech companies, and a handful of large research labs. As ML models are used with the final intention of supporting decision making for users, a great deal of information is needed in order to relate the user's decision to the solution given by the model. Deep learning models have shown promising performances in many tasks and domains when given a large enough dataset [33]. Therefore, we recommend collecting sufficiently large datasets to facilitate the use of deep learning models that could benefit risk detection systems in future studies. To overcome some of these data limitations, pre-trained deep learning models can also be used to address the lack of large labeled data which constrained some papers [39, 141, 147]. That said, given recent findings on the underlying biases that characterize pre-trained deep learning models, such as language models [28, 48], we suggest researchers to adopt ample caution when they are appropriated, such as considering debiasing approaches in concert [25]. Additionally, it could be interesting to explore data augmentation techniques to automatically supplement training data.

Another common shortcoming we observed is related to the classification setting of the task. The majority of the papers focused on binary classification, where the study would aim to identify an online risk instance or a predator. Binary classification fails to fully depict the characteristics of different sexual risks. Each sexual risk has its own set of stages and levels, which are hard to differentiate under the lens of binary classification. Future research should gear towards multi-class classifications based on the patterns of communication, focusing on different types of sexual risks as well as different stages and levels of each risk type. Multi-class classification will help researchers better understand sexual risks with respect to both categorical (types, levels) and temporal (stages) dimensions, which will be helpful in identifying online risk instances during the early stages, before the victim suffers from risk exposure for a prolonged period of time. Understanding the granularity level of analysis for conversations is important since approaches that focus on single messages would fail to understand the semantics and the context of conversations. Understanding individual messages is important when the context of the entire interaction is understood. Overall our results demonstrated that most approaches focused on identifying users, so future researchers need to develop approaches that take into account patterns and changes at the conversation level to detect the risks in earlier stages when it is happening.

Our finding that most studies used supervised algorithms implies that future works should develop more semi-supervised or a combination of supervised or unsupervised methods to provide automatic learning improvements through computational feedback. Particularly, unsupervised learning can address the challenges of not having big datasets and help provide insights about online risk. Unsupervised models can be considered as a first step used usually for understanding the underlying knowledge of the data to aid developers to create supervised models based on accurate understanding of the data. The prior understanding of the data is important not only for models' accuracy but also for models' interpretation by humans. In addition, human-in-the-loop

approaches and active learning (humans handling low confidence units and feeding those back into the model) [61] should be utilized so the models select what they need to learn next; that data may be sent to human annotators for training to teach edge cases, identify new categories, to help avoid over-fitting, or to help adapt to changing data characteristics and contexts.

6.1.6 Objective Performance Evaluation Are Not Enough. A common theme among the sexual detection approaches was that they came from a purely computational perspective and failed to incorporate any aspects of user-centered design or needs analysis. The articles did not include formative or summative user evaluations of the solutions developed. Instead, the majority of the papers reported computational evaluation metrics to demonstrate the performance of their risk detection models. The difference in classification tasks and the variety of evaluation metrics used in each study make it difficult to develop a standardized benchmark for computational evaluation. Standard metrics for measuring algorithmic performance for each specific risk detection task and guidelines for evaluating solutions' effectiveness and performance are needed for a direct comparison of different algorithms for each task. Although the current computational evaluation metrics are effective in the sense that they show computational performance, they could be further strengthened by assessments through user case studies and error analysis with examples. In addition to the algorithm evaluation metrics, it can be valuable to conduct user studies to evaluate the feasibility of the models in real-life scenarios. Tests and validations that involve humans should not only be done by the researchers but also by the actual stakeholders of the system for a more accurate evaluation [146].

Researchers have proposed different categories for evaluation methods or scales that could be adopted to ensure proper evaluation of the systems. For instance, Mohseni et al. [105] divided diverse types of evaluation methods into two groups including objective (task performance, user prediction of model output, compliance/reliance, etc.) and subjective measurements (interviews, surveys, self-reports, etc.) to be performed for evaluating the systems. They also categorized evaluation measures in the five themes Mental Models (how the AI works), Usefulness and Satisfaction, User Trust and Reliance, Human-AI Task Performance, and Computational Measures (correctness and completeness in terms of explaining what the model has learned). There exist some scales that can also be used by risk detection systems to evaluate their AI system, for instance, the "Explanation Satisfaction Scale" by Hoffman et al. [60] consisting of 8 items, which address factors such as understanding, satisfaction, completeness, accuracy, or trust. In addition, participatory design [17] could be used to involve people in the design of the algorithms to address the gaps between technical solutions and human expectations. Therefore, using different evaluation methods and scales as a checklist of items to be evaluated should be employed by researchers and reviewed in surveys.

6.1.7 Need for Artifacts and Embedding Models in Real-World Systems. We encourage future research to go beyond creating algorithms to developing technologies similar to how they would be used in practice. The majority of the studies fell short of implementing the algorithms in real-world systems or applications. Designs for these systems and interventions should be characterized to the relevant social groups and stakeholders such as adolescents, parents, police investigation teams, online moderators, social media companies, researchers, policymakers, practitioners, etc. from a Social Construction of Technology (SCOT) perspective which advocates that technology does not determine human action, but that rather, human action shapes technology [20]. Bringing user-centered perspectives to the forefront of sexual risk detection is necessary to evaluate the system by involving targeted users with the targeted context to make sure the artifacts meet the needs of different social user groups and the characteristics of the respective stakeholders. When creating ML algorithms, testing, and tuning the dataset is also important. It is important that future work leverage humans in the various circles of creating ML models including training, tuning, and

testing algorithms. At the same time, we do acknowledge that sometimes the choice of the ML model may limit the extent and utility of human involvement. For instance, deep learning models are inherently opaque and therefore model decisions or outcomes may not be easily understandable to laypersons in an intuitive manner. To support human involvement in such cases, researchers can consider approaches like "model cards" [102] to communicate to users about the limitations of the models and what the model outputs may mean.

From a practical standpoint, most articles that we reviewed did not declare the memory and the processing time of their approach, nor discuss the feasibility of implementing their model on mobile devices with fewer resources. Essentially as mobile devices are prevalent among youth which have their own specifications and limitations such as less computational power [148, 149]. Therefore to address the gap in the literature, researchers and practitioners need to design, implement and test approaches that work with mobile devices that considers computation, memory, and other limitations. Additionally, future researchers need to develop APIs that could be integrated into social media, chat rooms, forums, and other platforms for the detection of online sexual risks.

Furthermore, as we stated the gap of designing systems for respective stakeholders there are more points to consider when developing artifacts for different social user groups. These systems need to provide explainability of the AI system's functioning or decisions to be understandable by respected stakeholders [4]. These systems need to generate post-hoc explanations to justify an opaque model's decision in an accessible manner. Reasons for explanations include trustworthiness, causality, transferability, informativeness, fairness, accessibility, interactivity, or privacy awareness [166]. As the motivational, social, cognitive, along with professional and educational profiles of stakeholders affect their interpretations and reactions to explanations, so these explanations need to be suitable for each specific social group [4]. It is also important to explain why the stakeholders should trust the system. For instance, if the stakeholders are government agents or police officers, should certify the model compliance with the rules in force. Trustworthiness might be considered as the confidence of whether a model will act as intended when facing a given problem. There are cases in which the lack of a proper understanding of the model might drive the user toward incorrect assumptions and negative consequences. In this case, for instance, police officers or parents need to know exactly how much they can rely on the model, by receiving flagged content on what it means and how much they can trust the system. For example, a teen could send a picture with swimsuits that is detected as a sexual risk, but the parents should know how false positives might happen and know when to trust the detection system and when to manually check. These models should assess a generalization of robustness and stability, and confidence so humans can make the decisions on how to trust these systems.

6.2 The Importance of Human-Centeredness in Computational Risk Detection

6.2.1 Toward Victim and Survivor-Centered Sexual Risk Detection. To date, the primary focus has been detecting online sexual predators to help law enforcement agencies identify and prosecute sex offenders [21]. Although identifying predators is an important task, future work should consider identifying behaviors or indicators of being a victim of sexual risk. Victims are equally, if not the most, important stakeholders in the sexual risk incidents; examining how we can identify them could lead to strategies to support victims and even prevent any potential victims from going through such traumatic incidents. Identifying victims' reports can help concerned parties to pay attention to the violence reports and take reactions in a timely manner. Thus, the timing of the risk detection is of critical importance; the earlier we can detect risk exposure, the sooner we can intervene and mitigate the risks. Thus, artifacts and systems are needed to help prevent victimization rather than after-the-fact risk detection. Also, risk detection and mitigation come jointly; without a plan to mitigate risks, it would be useless to detect it. Therefore, when reviewing literature in

the area of risk detection, researchers need to take extra factors into consideration, for instance, the nature of the risk and the context in which it happens, the breakdown of the computational results and how that may affect the end-users, the timing of the detection, and the mitigation plan to protect the users. As HCI, HCML, and ML researchers, we are uniquely positioned, and thus arguably ethically obligated, to use our multi-disciplinary skills to not only accurately detect when sexual violence occurs online for the purpose of understanding this modern-day phenomenon, but to also become activists that work to eradicate sexually-motivated online violence.

6.2.2 Toward Human-Centeredness in Computational Risk Detection. Without applying the human-centered lens to the context of sexual risk detection, many of our insights would have been overlooked. We came to these insights about the gaps and opportunities in the literature by having the target users' needs in the valid real-world scenarios in mind. Many authors of the papers we reviewed did not mention the limitations of their research that we surfaced in this research. Some of these insights include the after-the-fact risk detection, lack of ecologically valid datasets, and lack of user studies and real-world evaluations. Overall, there were many valuable contributions made by the existing literature that should be continued in future work. For example, researchers from social sciences [15, 113] proposed theoretical frameworks for online grooming processes to help the developers of predators automated detection systems, leveraging the behavioral pattern recognition to improve educational tools for the community. For instance, Cycle of Entrapment from the Luring Communication Theory (LCT) [110] was developed by social scientists related to the different stages of how sexual grooming happens. This theory has/can be used to refine the data annotation and feature selection processes for AI model development to improve accuracy through insight into human behavior. This would contrast with solely data-driven approaches. Human knowledge also includes understanding social and human biases and trying to adjust that for algorithms to balance unfair decisions. The increase in human theory and involvement increases the interpretability and the practicality of the models, as these models, in the end, serve the sole purpose to aid and help the people. Therefore, we need increased but also well-monitored and well-reformed involvement of humans in AI risk detection design. Similarly, we urge computational ML experts to work with HCI researchers and social scientists when dealing with real-world human problems to incorporate human and social interpretations in the design and development of the models. If we want to protect people from online risks we would need a close collaboration with policymakers, psychologists, and lawmakers.

Computational online risk detection, in the end, is a problem that originates between the online interactions between people and aims to protect people and mitigate any potential risks. Thus, our human-centered lens can also be applicable to other domains of online risk detection, such as cyberbullying or hate speech. Therefore, other researchers could use our human-centered lens for reviewing the computational risk detection literature beyond that of online sexual risk detection. Moreover, it is important that researchers and practitioners respond to the call for the need to add FATE (fairness, accountability, transparency, and Ethics) from the point of view of a human observer to explain the reasons behind decisions or the processes that generate them [10] which is missing from the current sexual risk detection literature.

6.3 Moving Beyond Human-Centeredness to Examine and Embed Values in HCML

According to Adadi et al. [4] the need for explaining AI systems comes from four motivations. First, there is the need to ensure that AI-based decisions were not made erroneously with justifications for each outcome. Second, explanations make vulnerabilities and flaws visible so that they could be controlled and corrected. Third, explainability is needed so others can continuously improve the models with a better understanding of the capabilities and implications of the models. Fourth,

providing explanations could potentially lead to discovering new facts and gaining more knowledge. It is important to consider AI systems' fairness to avoid "unfair" algorithms/systems whose decisions are skewed toward a particular group of people. According to Mehrabi et al.'s [96] survey of different biases in ML application, there are two sources of unfairness including those coming from biases in data and those stemming from algorithms. ML researchers need to use such taxonomies for fairness to avoid biases in the technologies that they develop.

Therefore, researchers need to think more about ethics and speculate more on how these models could be used by bad actors. There are ways that sexual risk detection models could be used in adverse ways to harm people instead of helping them [129]. As an example, sexual predators could use trained computer vision algorithms to find child pornography. So it is important to make sure that proper safeguards are in place to protect people. In order to do so, we advocate that studies can take a speculative approach [17] on how these algorithms might be used to harm. Moreover, privacy awareness of the model by users is important since models may have complex representations of their learned patterns and not being able to understand what has been captured by the model and stored in its internal representation may entail a privacy breach Mosallanezhad et al. [107]. Thus for researchers and practitioners that develop these types of systems need to think about potential privacy breaches of users and potential misuses of the ML model [126]. There was a universal lack of any explicit mention of the ethics, potential biases or speculations on the usages of the models in the reviewed papers. Exceptions include Chowdhury et al. [49], who discussed privacy of individuals, fairness, bias, discrimination, and interpretation of their study. The literature being scant on FATE topics was our barrier to qualitatively coding the papers for these values in our literature review in a systematic way. Future studies should move beyond considering only computational approaches and consider more human values embedded in their research, following the recommendation of Jo and Gebru [72] around requiring and building suitable institutional frameworks and procedures.

7 CONCLUSION

We reviewed 73 studies on computational approaches to online sexual risk detection utilizing a human-centered lens, which may be used by researchers for reviewing computational risk detection research more broadly. Our review provided insights into the trends and gaps within the current literature and opportunities for future research. Although our literature review is on computational methods, as sexual risk detection is a social issue that needs integration of personal, social, and cultural aspects for designing and developing intelligent systems that understand this socially complicated issue, reviewing literature from a human perspective is necessary. We call on the community to design and develop human-centered solutions to address these gaps by considering the stakeholders at all stages of designing and developing technologies that bridge the socio-technical gaps for sexual risk detection.

ACKNOWLEDGMENTS

We thank Shiza Ali who assisted with the articles qualitative coding. This research is partially supported by the U.S. National Science Foundation under grants #IIP-1827700 and #IIS-1844881 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.

REFERENCES

[1] 2012. Fact Sheet: the Obama Administration Announces Efforts to Combat Human Trafficking at Home and Abroad. https://obamawhitehouse.archives.gov/the-press-office/2012/09/25/fact-sheet-obama-administration-

announces-efforts-combat-human-traffickі

[2] 2020. *National Center for Missing and Exploited Children.* https://www.missingkids.org/footer/media/keyfacts

[3] Trafficking Victims Protection Act. 2000. Victims of trafficking and violence protection act of 2000. *United States* (2000).

[4] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE access* 6 (2018), 52138–52160.

[5] Hala Al Kuwatly, Maximilian Wich, and Georg Groh. 2020. Identifying and measuring annotator bias based on annotators' demographic characteristics. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*. 184–190.

[6] Ali Alkhatib and Michael Bernstein. 2019. Street-level algorithms: A theory at the gaps between policy and decisions. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.

[7] Faith Amuchi, Ameer Al-Nemrat, Mamoun Alazab, and Robert Layton. 2012. Identifying cyber predators through forensic authorship analysis of chat logs. In *2012 Third Cybercrime and Trustworthy Computing Workshop*. IEEE, 28–37.

[8] 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 *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16*. ACM Press, Santa Clara, California, USA, 3906–3918. https://doi.org/10.1145/2858036.2858096

[9] Philip Anderson, Zheming Zuo, Longzhi Yang, and Yanpeng Qu. 2019. An Intelligent Online Grooming Detection System Using AI Technologies. In *2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. 1–6. https://doi.org/10.1109/FUZZ-IEEE.2019.8858973 ISSN: 1544-5615.

[10] 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.

[11] 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

[12] Karla Badillo-Urquiola, Afsaneh Razi, Jan Edwards, and Pamela Wisniewski. 2020. Children's Perspectives on Human Sex Trafficking Prevention Education. In *Companion of the 2020 ACM International Conference on Supporting Group Work*. 123–126.

[13] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence* 41, 2 (2018), 423–443.

[14] Rahaf Barakat, Sameer Abufardeh, and Kenneth Magel. 2016. Automated framework to improve users' awareness on Online Social Networks. In *2016 IEEE International Conference on Electro Information Technology (EIT)*. IEEE, 0428–0433.

[15] Connie Barber and S.C. Bettez. 2014. Deconstructing the online grooming of youth: Toward improved information systems for detection of online sexual predators. (Jan. 2014).

[16] Ravi Barreira, Vládia Pinheiro, and Vasco Furtado. 2017. A framework for digital forensics analysis based on semantic role labeling. In *2017 IEEE International Conference on Intelligence and Security Informatics (ISI)*. IEEE, 66–71.

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

[18] Victoria Bellotti and Keith Edwards. 2001. Intelligibility and accountability: human considerations in context-aware systems. *Human–Computer Interaction* 16, 2-4 (2001), 193–212.

[19] Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics* 6 (2018), 587–604.

[20] Wiebe E Bijker, Thomas P Hughes, Trevor Pinch, et al. 1987. The social construction of technological systems.

[21] Pamela J. Black, Melissa Wollis, Michael Woodworth, and Jeffrey T. Hancock. 2015. A linguistic analysis of grooming strategies of online child sex offenders: Implications for our understanding of predatory sexual behavior in an increasingly computer-mediated world. *Child Abuse & Neglect* 44 (June 2015), 140–149. https://doi.org/10.1016/j.chiabu.2014.12.004

[22] Dasha Bogdanova, Paolo Rosso, and Thamar Solorio. 2012. Modelling Fixated Discourse in Chats with Cyberpedophiles. In *Proceedings of the Workshop on Computational Approaches to Deception Detection (EACL 2012)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 86–90. http://dl.acm.org/citation.cfm?id=2388616.2388629 event-place: Avignon, France.

[23] Dasha Bogdanova, Paolo Rosso, and Thamar Solorio. 2012. On the Impact of Sentiment and Emotion Based Features in Detecting Online Sexual Predators. In *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis (WASSA '12)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 110–118.

http://dl.acm.org/citation.cfm?id=2392963.2392986

[24] Dasha Bogdanova, Paolo Rosso, and Thamar Solorio. 2014. Exploring high-level features for detecting cyberpedophilia. *Computer speech & language* 28, 1 (2014), 108–120.

[25] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in Neural Information Processing Systems* (2016).

[26] Parisa Rezaee Borj and Patrick Bours. 2019. Predatory Conversation Detection. In *2019 International Conference on Cyber Security for Emerging Technologies (CSET)*. 1–6. https://doi.org/10.1109/CSET.2019.8904885 ISSN: null.

[27] 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.

[28] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. PMLR, 77–91.

[29] Ángel Callejas-Rodríguez, Esaú Villatoro-Tello, Ivan Meza, and Gabriela Ramírez-de-la Rosa. 2016. From Dialogue Corpora to Dialogue Systems: Generating a Chatbot with Teenager Personality for Preventing Cyber-Pedophilia. In *Text, Speech, and Dialogue*, Petr Sojka, Aleš Horák, Ivan Kopeček, and Karel Pala (Eds.). Vol. 9924. Springer International Publishing, Cham, 531–539. https://doi.org/10.1007/978-3-319-45510-5_61

[30] Amparo Elizabeth Cano, Miriam Fernandez, and Harith Alani. 2014. Detecting Child Grooming Behaviour Patterns on Social Media. In *Social Informatics: 6th International Conference, SocInfo 2014, Barcelona, Spain, November 11-13, 2014. Proceedings*, Luca Maria Aiello and Daniel McFarland (Eds.). Springer International Publishing, Cham, 412–427. https://doi.org/10.1007/978-3-319-13734-6_30

[31] Stevie Chancellor, Eric P. S. Baumer, and Munmun De Choudhury. 2019. Who is the "Human" in Human-Centered Machine Learning: The Case of Predicting Mental Health from Social Media. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 147 (Nov. 2019), 32 pages. https://doi.org/10.1145/3359249

[32] Jonathan H Chen and Abraham Verghese. 2020. Planning for the Known Unknown: Machine Learning for Human Healthcare Systems. *The American Journal of Bioethics* 20, 11 (2020), 1–3.

[33] Xue-Wen Chen and Xiaotong Lin. 2014. Big data deep learning: challenges and perspectives. *IEEE access* 2 (2014), 514–525.

[34] Lu Cheng, Ahmadreza Mosallanezhad, Yasin Silva, Deborah Hall, and Huan Liu. 2021. Mitigating Bias in Session-based Cyberbullying Detection: A Non-Compromising Approach. In *Proceedings of ACL*.

[35] Y. Cheong, A. K. Jensen, E. R. Guðnadóttir, B. Bae, and J. Togelius. 2015. Detecting Predatory Behavior in Game Chats. *IEEE Transactions on Computational Intelligence and AI in Games* 7, 3 (Sept. 2015), 220–232. https://doi.org/10.1109/TCIAIG.2015.2424932

[36] Ming Ming Chiu, Kathryn C Seigfried-Spellar, and Tatiana R Ringenberg. 2018. Exploring detection of contact vs. fantasy online sexual offenders in chats with minors: Statistical discourse analysis of self-disclosure and emotion words. *Child abuse & neglect* 81 (2018), 128–138.

[37] Zeineb Dhouioui and Jalel Akaichi. 2016. Privacy Protection Protocol in Social Networks Based on Sexual Predators Detection. In *Proceedings of the International Conference on Internet of Things and Cloud Computing (ICC '16)*. ACM, New York, NY, USA, 63:1–63:6. https://doi.org/10.1145/2896387.2896448

[38] Mohammadreza Ebrahimi and Mohammadreza Ebrahimi. 2016. *Automatic Identification of Online Predators in Chat Logs by Anomaly Detection and Deep Learning*. Master's thesis. Concordia University. https://spectrum.library.concordia.ca/981404/

[39] Mohammadreza Ebrahimi, Ching Y. Suen, and Olga Ormandjieva. 2016. Detecting predatory conversations in social media by deep Convolutional Neural Networks. *Digital Investigation* 18 (Sept. 2016), 33–49. https://doi.org/10.1016/j.diin.2016.07.001

[40] Mohammadreza Ebrahimi, Ching Y. Suen, Olga Ormandjieva, and Adam Krzyzak. 2016. Recognizing Predatory Chat Documents using Semi-supervised Anomaly Detection. https://www.ingentaconnect.com/content/ist/ei/2016/00002016/00000017/art00012#

[41] Paul Elzinga, Karl Erich Wolff, and Jonas Poelmans. 2012. Analyzing Chat Conversations of Pedophiles with Temporal Relational Semantic Systems. In *2012 European Intelligence and Security Informatics Conference*. 242–249. https://doi.org/10.1109/EISIC.2012.12

[42] Hugo Jair Escalante, Esaú Villatoro-Tello, Antonio Juárez, Manuel Montes-y Gómez, and Luis Villaseñor. 2013. Sexual predator detection in chats with chained classifiers. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Association for Computational Linguistics, Atlanta, Georgia, 46–54. http://www.aclweb.org/anthology/W13-1607

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

9107945

[44] Christiane Fellbaum. 2010. WordNet. In *Theory and applications of ontology: computer applications*. Springer, 231–243.

[45] Joseph L Fleiss and Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement* 33, 3 (1973), 613–619.

[46] R. Fournier and M. Danisch. 2014. Mining bipartite graphs to improve semantic pedophile activity detection. In *2014 IEEE Eighth International Conference on Research Challenges in Information Science (RCIS)*. 1–4. https://doi.org/10.1109/RCIS.2014.6861035

[47] Fencepost Frank and Von Erck Xavier. [n.d.]. Perverted-Justice.com - The largest and best anti-predator organization online. http://www.perverted-justice.com/

[48] Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences* 115, 16 (2018), E3635–E3644.

[49] 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 *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*. Association for Computational Linguistics, Minneapolis, Minnesota, 136–146. https://doi.org/10.18653/v1/N19-3018

[50] L. Gillam and A. Vartapetiance. 2012. Quite Simple Approaches for Authorship Attribution, Intrinsic Plagiarism Detection and Sexual Predator Identification. In *LNCS*. Rome, Italy. http://www.uni-weimar.de/medien/webis/research/events/pan-12/pan12-web/

[51] Fergyanto E. Gunawan, Livia Ashianti, Sevenpri Candra, and Benfano Soewito. 2016. Detecting online child grooming conversation. In *2016 11th International Conference on Knowledge, Information and Creativity Support Systems (KICSS)*. 1–6. https://doi.org/10.1109/KICSS.2016.7951413

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

[53] Rex Hartson and Pardha Pyla. 2019. Chapter 22 - Empirical UX Evaluation: UX Goals, Metrics, and Targets. In *The UX Book (Second Edition)* (second edition ed.), Rex Hartson and Pardha Pyla (Eds.). Morgan Kaufmann, Boston, 453–481. https://doi.org/10.1016/B978-0-12-805342-3.00022-9

[54] 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.

[55] Feng He, Yu Deng, and Weina Li. 2020. Coronavirus disease 2019: What we know? *Journal of medical virology* 92, 7 (2020), 719–725.

[56] Nicola Henry and Anastasia Powell. 2018. Technology-facilitated sexual violence: A literature review of empirical research. *Trauma, violence, & abuse* 19, 2 (2018), 195–208.

[57] José María Gómez Hidalgo and Andrés Alfonso Caurcel Díaz. [n.d.]. Combining Predation Heuristics and Chat-Like Features in Sexual Predator Identification. ([n. d.]), 6.

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

[59] 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.

[60] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608* (2018).

[61] Andreas Holzinger. 2016. Interactive machine learning for health informatics: when do we need the human-in-the-loop? *Brain Informatics* 3, 2 (2016), 119–131.

[62] Aziz Z Huq. 2018. Racial equity in algorithmic criminal justice. *Duke LJ* 68 (2018), 1043.

[63] M. Ibanez and R. Gazan. 2016. Detecting sex trafficking circuits in the U.S. through analysis of online escort advertisements. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. 892–895. https://doi.org/10.1109/ASONAM.2016.7752344

[64] Michelle Ibanez and Rich Gazan. 2016. Virtual indicators of sex trafficking to identify potential victims in online advertisements. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 818–824.

[65] Michelle Ibanez and Daniel D Suthers. 2014. Detection of domestic human trafficking indicators and movement trends using content available on open internet sources. In *2014 47th Hawaii international conference on system sciences*. IEEE, 1556–1565.

[66] M. Ibanez and D. D. Suthers. 2016. Detecting covert sex trafficking networks in virtual markets. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. 876–879. https://doi.org/10.1109/ASONAM.2016.7752340

[67] 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.

[68] Giacomo Inches, Morgan Harvey, and Fabio Crestani. 2013. Finding participants in a chat: Authorship attribution for conversational documents. In *2013 International Conference on Social Computing*. IEEE, 272–279.

[69] F. Iqbal, B. C. M. Fung, M. Debbabi, R. Batool, and A. Marrington. 2019. Wordnet-Based Criminal Networks Mining for Cybercrime Investigation. *IEEE Access* 7 (2019), 22740–22755. https://doi.org/10.1109/ACCESS.2019.2891694

[70] Shunichi Ishihara. 2014. A comparative study of likelihood ratio based forensic text comparison procedures: Multivariate kernel density with lexical features vs. word N-grams vs. character N-grams. In *2014 Fifth Cybercrime and Trustworthy Computing Conference*. IEEE, 1–11.

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

[72] Eun Seo Jo and Timnit Gebru. 2020. Lessons from archives: Strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 306–316.

[73] 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

[74] Michael Katell, Meg Young, Dharma Dailey, Bernease Herman, Vivian Guetler, Aaron Tam, Corinne Bintz, Daniella Raz, and PM Krafft. 2020. Toward situated interventions for algorithmic equity: lessons from the field. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 45–55.

[75] Aparup Khatua, Erik Cambria, and Apalak Khatua. 2018. Sounds of Silence Breakers: Exploring Sexual Violence on Twitter. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. 397–400. https://doi.org/10.1109/ASONAM.2018.8508576 ISSN: 2473-9928.

[76] Jinhwa Kim, Yoon Jo Kim, Mitra Behzadi, and Ian G. Harris. 2020. Analysis of Online Conversations to Detect Cyberpredators Using Recurrent Neural Networks. In *Proceedings for the First International Workshop on Social Threats in Online Conversations: Understanding and Management*. European Language Resources Association, Marseille, France, 15–20. https://www.aclweb.org/anthology/2020.stoc-1.3

[77] 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 *Proceedings of the International AAAI Conference on Web and Social Media*.

[78] Rob Kling and Susan Leigh Star. 1998. Human Centered Systems in the Perspective of Organizational and Social Informatics. *SIGCAS Comput. Soc.* 28, 1 (March 1998), 22–29. https://doi.org/10.1145/277351.277356

[79] April Kontostathis. 2009. ChatCoder: Toward the Tracking and Categorization of Internet Predators. In *Proc. Text Mining Workshop 2009 Held in Conjunction with the Ninth Siam International Conference on Data Mining (sdm 2009). Sparks, Nv. May 2009*.

[80] April Kontostathis, Lynne Edwards, Jen Bayzick, Amanda Leatherman, and Kristina Moore. 2009. Comparison of rule-based to human analysis of chat logs. *communication theory* 8, 2 (2009).

[81] Panos Kostakos, Lucie Špráchalová, Abhinay Pandya, Mohamed Aboeleinen, and Mourad Oussalah. 2018. Covert Online Ethnography and Machine Learning for Detecting Individuals at Risk of Being Drawn into Online Sex Work. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. 1096–1099. https://doi.org/10.1109/ASONAM.2018.8508276 ISSN: 2473-991X.

[82] Michelle A Krieger. 2017. Unpacking "sexting": A systematic review of nonconsensual sexting in legal, educational, and psychological literatures. *Trauma, Violence, & Abuse* 18, 5 (2017), 593–601.

[83] Scott Krig. 2016. Ground truth data, content, metrics, and analysis. In *Computer Vision Metrics*. Springer, 247–271.

[84] Carlos Laorden, Patxi Galán-García, Igor Santos, Borja Sanz, Jose Gomez Hidalgo, and Pablo Bringas. 2013. Negobot: A Conversational Agent Based on Game Theory for the Detection of Paedophile Behaviour. In *Advances in Intelligent Systems and Computing*. Vol. 189. 261–270. https://doi.org/10.1007/978-3-642-33018-6_27

[85] Matthieu Latapy, Clémence Magnien, and Raphaël Fournier. 2011. Quantifying paedophile queries in a large P2P system. In *2011 Proceedings IEEE INFOCOM*. 401–405. https://doi.org/10.1109/INFCOM.2011.5935191 ISSN: 0743-166X.

[86] L. Li, O. Simek, A. Lai, M. Daggett, C. K. Dagli, and C. Jones. 2018. Detection and Characterization of Human Trafficking Networks Using Unsupervised Scalable Text Template Matching. In *2018 IEEE International Conference on Big Data (Big Data)*. 3111–3120. https://doi.org/10.1109/BigData.2018.8622189

[87] Richard J Light. 1971. Measures of response agreement for qualitative data: some generalizations and alternatives. *Psychological bulletin* 76, 5 (1971), 365.

[88] Yingchi Liu, Quanzhi Li, Xiaozhong Liu, Qiong Zhang, and Luo Si. 2019. Sexual Harassment Story Classification and Key Information Identification. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM '19)*. Association for Computing Machinery, Beijing, China, 2385–2388. https://doi.org/10.1145/3357384.3358146

[89] Sonia Livingstone and Leslie Haddon. 2008. Risky experiences for children online: charting European research on children and the Internet. *Children and Society* 22 (July 2008), 314–323. http://www.wiley.com/bw/journal.asp?ref=0951-0605

[90] 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 (Jan. 2020), 15–27. https://doi.org/10.1016/j.pragma.2019.09.010

[91] Adrian Pastor López-Monroy, Fabio A. González, Manuel Montes, Hugo Jair Escalante, and Thamar Solorio. 2018. Early Text Classification Using Multi-Resolution Concept Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 1216–1225. https://doi.org/10.18653/v1/N18-1110

[92] Katrinna MacFarlane and Violeta Holmes. 2009. Agent-Mediated Information Exchange: Child Safety Online. In *2009 International Conference on Management and Service Science*. 1–5. https://doi.org/10.1109/ICMSS.2009.5302027 ISSN: null.

[93] Ann Majchrzak, Samer Faraj, Gerald C Kane, and Bijan Azad. 2013. The contradictory influence of social media affordances on online communal knowledge sharing. *Journal of Computer-Mediated Communication* 19, 1 (2013), 38–55.

[94] Nora McDonald, Karla Badillo-Urquiola, Morgan G Ames, Nicola Dell, Elizabeth Keneski, Manya Sleeper, and Pamela J Wisniewski. 2020. Privacy and Power: Acknowledging the Importance of Privacy Research and Design for Vulnerable Populations. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–8.

[95] India McGhee, Jennifer Bayzick, April Kontostathis, Lynne Edwards, Alexandra McBride, and Emma Jakubowski. 2011. Learning to Identify Internet Sexual Predation. *International Journal of Electronic Commerce* 15, 3 (April 2011), 103–122. https://doi.org/10.2753/JEC1086-4415150305

[96] 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).

[97] Md. Waliur Rahman Miah, John Yearwood, and Sid Kulkarni. 2011. Detection of child exploiting chats from a mixed chat dataset as a text classification task. In *Proceedings of the Australasian Language Technology Association Workshop 2011*. Canberra, Australia, 157–165. https://www.aclweb.org/anthology/U11-1020

[98] D. Michalopoulos and I. Mavridis. 2011. Utilizing document classification for grooming attack recognition. In *2011 IEEE Symposium on Computers and Communications (ISCC)*. 864–869. https://doi.org/10.1109/ISCC.2011.5983950

[99] D. Michalopoulos, E. Papadopoulos, and I. Mavridis. 2012. Artemis: Protection from Sexual Exploitation Attacks via SMS. In *2012 16th Panhellenic Conference on Informatics*. 19–24. https://doi.org/10.1109/PCi.2012.46

[100] Pushkar Mishra, Helen Yannakoudakis, and Ekaterina Shutova. 2019. Tackling online abuse: A survey of automated abuse detection methods. *arXiv preprint arXiv:1908.06024* (2019).

[101] Kanishka Misra, Hemanth Devarapalli, Tatiana R. Ringenberg, and Julia Taylor Rayz. 2019. Authorship Analysis of Online Predatory Conversations using Character Level Convolution Neural Networks. In *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*. 623–628. https://doi.org/10.1109/SMC.2019.8914323 ISSN: 1062-922X.

[102] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*. 220–229.

[103] Miljana Mladenović, Vera Ošmjanski, and Staša Vujičić Stanković. 2021. Cyber-aggression, Cyberbullying, and Cyber-grooming: A Survey and Research Challenges. *ACM Computing Surveys (CSUR)* 54, 1 (2021), 1–42.

[104] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. 2018. *Foundations of machine learning*. MIT press.

[105] Sina Mohseni, Niloofar Zarei, and Eric D Ragan. 2018. A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *arXiv preprint arXiv:1811.11839* (2018).

[106] Camille Mori, Jessica E Cooke, Jeff R Temple, Anh Ly, Yu Lu, Nina Anderson, Christina Rash, and Sheri Madigan. 2020. The prevalence of sexting behaviors among emerging adults: A meta-analysis. *Archives of sexual behavior* (2020), 1–17.

[107] Ahmadreza Mosallanezhad, Ghazaleh Beigi, and Huan Liu. 2019. Deep reinforcement learning-based text anonymization against private-attribute inference. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 2360–2369.

[108] C. H. Ngejane, G. Mabuza-Hocquet, J. H. P. Eloff, and S. Lefophane. 2018. Mitigating Online Sexual Grooming Cybercrime on Social Media Using Machine Learning: A Desktop Survey. In *2018 International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD)*. 1–6. https://doi.org/10.1109/ICABCD.2018.8465413

[109] Giang Hoang Nguyen, Abdesselam Bouzerdoum, and Son Lam Phung. 2009. Learning pattern classification tasks with imbalanced data sets. *Pattern recognition* (2009), 193–208.

[110] Loreen N. Olson, Joy L. Daggs, Barbara L. Ellevold, and Teddy K. K. Rogers. 2007. Entrapping the Innocent: Toward a Theory of Child Sexual Predators' Luring Communication. *Communication Theory* 17, 3 (2007), 231–251. https://doi.org/10.1111/j.1468-2885.2007.00294.x arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1468-2885.2007.00294.x

[111] Cathy O'neil. 2016. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.

[112] Esteban Ortiz-Ospina. [n.d.]. *The rise of social media*. https://ourworldindata.org/rise-of-social-media

[113] Rachel O'Connell. 2003. A typology of child cybersexploitation and online grooming practices. *Cyberspace Research Unit, University of Central Lancashire* (2003).

[114] Alexander Panchenko, Richard Beaufort, Hubert Naets, and Cédrick Fairon. 2013. Towards Detection of Child Sexual Abuse Media: Categorization of the Associated Filenames. In *Advances in Information Retrieval (Lecture Notes in Computer Science)*, Pavel Serdyukov, Pavel Braslavski, Sergei O. Kuznetsov, Jaap Kamps, Stefan Rüger, Eugene Agichtein, Ilya Segalovich, and Emine Yilmaz (Eds.). Springer Berlin Heidelberg, 776–779.

[115] Suraj Jung Pandey, Ioannis Klapaftis, and Suresh Manandhar. 2012. Detecting Predatory Behaviour from Online Textual Chats. In *Multimedia Communications, Services and Security (Communications in Computer and Information Science)*, Andrzej Dziech and Andrzej Czyżewski (Eds.). Springer, Berlin, Heidelberg, 270–281. https://doi.org/10.1007/978-3-642-30721-8_27

[116] Claudia Peersman, Frederik Vaassen, and Vincent Van Asch. 2012. Conversation Level Constraints on Pedophile Detection in Chat Rooms. (2012), 13.

[117] N. Pendar. 2007. Toward Spotting the Pedophile Telling victim from predator in text chats. In *International Conference on Semantic Computing (ICSC 2007)*. 235–241. https://doi.org/10.1109/ICSC.2007.32

[118] L. Penna, A. Clark, and G. Mohay. 2010. A Framework for Improved Adolescent and Child Safety in MMOs. In *2010 International Conference on Advances in Social Networks Analysis and Mining*. 33–40. https://doi.org/10.1109/ASONAM.2010.66

[119] 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.

[120] Ralph M Perhac Jr. 1996. Defining risk: Normative considerations. *Human and Ecological Risk Assessment* 2, 2 (1996), 381–392.

[121] Anthony T. Pinter, Pamela J. Wisniewski, Heng Xu, Mary Beth Rosson, and Jack M. Caroll. 2017. Adolescent Online Safety: Moving Beyond Formative Evaluations to Designing Solutions for the Future. In *Proceedings of the 2017 Conference on Interaction Design and Children - IDC '17*. ACM Press, Stanford, California, USA, 352–357. https://doi.org/10.1145/3078072.3079722

[122] Kate Nalini Murbade Arun Pokharkar Anuja, Shelake Shubham. 2015. Protective shield for social networks to defend cyberbullying and online grooming attacks. In *International Journal of Advances in Electronics and Computer Science*, Vol. 2.

[123] N. Potha, M. Maragoudakis, and D. Lyras. 2016. A biology-inspired, data mining framework for extracting patterns in sexual cyberbullying data. *Knowledge-Based Systems* 96 (March 2016), 134–155. https://doi.org/10.1016/j.knosys.2015.12.021

[124] Hady Pranoto, Fergyanto E Gunawan, and Benfano Soewito. 2015. Logistic models for classifying online grooming conversation. *Procedia Computer Science* 59 (2015), 357–365.

[125] Francisco Rangel, Paolo Rosso, Moshe Koppel, Efstathios Stamatatos, and Giacomo Inches. 2013. Overview of the author profiling task at pan 2013. In *CLEF Conference on Multilingual and Multimodal Information Access Evaluation*. CELCT, 352–365.

[126] 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*.

[127] Afsaneh Razi, Karla Badillo-Urquiola, and Pamela J. Wisniewski. 2020. Let's Talk about Sext: How Adolescents Seek Support and Advice about Their Online Sexual Experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, Honolulu, HI, USA, 1–13. https://doi.org/10.1145/3313831.3376400

[128] Afsaneh Razi, Seunghyun Kim, Ashwaq Alsoubai, Xavier Caddle, Shiza Ali, Munmun De Choudhury, Pamela Wisniewski, et al. 2021. Teens at the Margin: Artificially Intelligent Technology for Promoting Adolescent Online Safety. In *ACM Conference on Human Factors in Computing Systems (CHI 2021)/Artificially Intelligent Technology for the Margins: A Multidisciplinary Design Agenda Workshop*.

[129] Afsaneh Razi, Seunghyun Kim, Munmun De Choudhury, and Pamela Wisniewski. 2019. Ethical considerations for adolescent online risk detection AI systems. In *Good Systems: Ethical AI for CSCW (The 22nd ACM Conference on Computer-Supported Cooperative Work and Social Computing)*.

[130] T. Ringenberg, K. Misra, K. C. Seigfried-Spellar, and J. Taylor Rayz. 2019. Exploring Automatic Identification of Fantasy-Driven and Contact-Driven Sexual Solicitors. In *2019 Third IEEE International Conference on Robotic Computing (IRC)*. 532–537. https://doi.org/10.1109/IRC.2019.00110

[131] 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.

[132] Aja Romano. [n.d.]. *A new law intended to curb sex trafficking threatens the future of the internet as we know it.* https://www.vox.com/culture/2018/4/13/17172762/fosta-sesta-backpage-230-internet-freedom

[133] Hugo Rosa, David Matos, Ricardo Ribeiro, Luisa Coheur, and João P Carvalho. 2018. A "deeper" look at detecting cyberbullying in social networks. In *2018 international joint conference on neural networks (IJCNN)*. IEEE, 1–8.

[134] Michael J Rosenfeld, Reuben J Thomas, and Sonia Hausen. 2019. Disintermediating your friends: How online dating in the United States displaces other ways of meeting. *Proceedings of the National Academy of Sciences* 116, 36 (2019), 17753–17758.

[135] Mattias Rost, Louise Barkhuus, Henriette Cramer, and Barry Brown. 2013. Representation and communication: challenges in interpreting large social media datasets. (2013), 6.

[136] T. Roy, J. McClendon, and L. Hodges. 2018. Analyzing Abusive Text Messages to Detect Digital Dating Abuse. In *2018 IEEE International Conference on Healthcare Informatics (ICHI)*. 284–293. https://doi.org/10.1109/ICHI.2018.00039

[137] Moshe Rutgaizer, Yuval Shavitt, Omer Vertman, and Noa Zilberman. 2012. Detecting Pedophile Activity in BitTorrent Networks. In *Passive and Active Measurement (Lecture Notes in Computer Science)*, Nina Taft and Fabio Ricciato (Eds.). Springer, Berlin, Heidelberg, 106–115. https://doi.org/10.1007/978-3-642-28537-0_11

[138] Dominik Sacha, Michael Sedlmair, Leishi Zhang, John Aldo Lee, Daniel Weiskopf, Stephen North, and Daniel Keim. 2016. Human-centered machine learning through interactive visualization. In *ESANN 2016: 24th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning Bruges, Belgium April 27-28-29, 2016 Proceedings*. ESANN, Bruges, Belgium, 641–646. https://www.elen.ucl.ac.be/Proceedings/esann/esannpdf/es2016-166.pdf

[139] Devansh Saxena, Karla Badillo-Urquiola, Pamela J Wisniewski, and Shion Guha. 2020. A Human-Centered Review of Algorithms used within the US Child Welfare System. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–15.

[140] 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

[141] Saeideh Shahrokh Esfahani, Michael J. Cafarella, Maziyar Baran Pouyan, Gregory DeAngelo, Elena Eneva, and Andy E. Fano. 2019. Context-specific Language Modeling for Human Trafficking Detection from Online Advertisements. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 1180–1184. https://doi.org/10.18653/v1/P19-1114

[142] Daniel Ribeiro Silva, Andrew Philpot, Abhishek Sundararajan, Nicole Marie Bryan, and Eduard Hovy. 2014. Data integration from open internet sources and network detection to combat underage sex trafficking. In *Proceedings of the 15th Annual International Conference on Digital Government Research*. 86–90.

[143] Thamar Solorio, Mahsa Shafaei, Christos Smailis, Isabelle Augenstein, Margaret Mitchell, Ingrid Stapf, and Ioannis Kakadiaris. [n.d.]. White Paper-Creating a Repository of Objectionable Online Content: Addressing Undesirable Biases and Ethical Considerations. ([n. d.]).

[144] Brian H Spitzberg and Gregory Hoobler. 2002. Cyberstalking and the technologies of interpersonal terrorism. *New media & society* 4, 1 (2002), 71–92.

[145] Sudha Subramani, Hua Wang, Md Rafiqul Islam, Anwaar Ulhaq, and Manjula O'Connor. 2018. Child Abuse and Domestic Abuse: Content and Feature Analysis from Social Media Disclosures. In *Databases Theory and Applications (Lecture Notes in Computer Science)*, Junhu Wang, Gao Cong, Jinjun Chen, and Jianzhong Qi (Eds.). Springer International Publishing, Cham, 174–185. https://doi.org/10.1007/978-3-319-92013-9_14

[146] Harini Suresh, Steven R Gomez, Kevin K Nam, and Arvind Satyanarayan. 2021. Beyond Expertise and Roles: A Framework to Characterize the Stakeholders of Interpretable Machine Learning and their Needs. *arXiv preprint arXiv:2101.09824* (2021).

[147] Ashima Suvarna, Grusha Bhalla, Shailender Kumar, and Ashi Bhardwaj. 2020. Identifying Victim Blaming Language in Discussions about Sexual Assaults on Twitter. In *International Conference on Social Media and Society (SMSociety'20)*. Association for Computing Machinery, Toronto, ON, Canada, 156–163. https://doi.org/10.1145/3400806.3400825

[148] M. U. Tariq, A. K. Ghosh, K. Badillo-Urquiola, A. Jha, S. Koppal, and P. J. Wisniewski. 2018. Designing Light Filters to Detect Skin Using a Low-powered Sensor. In *SoutheastCon 2018*. 1–8. https://doi.org/10.1109/SECON.2018.8479027

[149] 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

[150] Fujio Toriumi, Takafumi Nakanishi, Mitsuteru Tashiro, and Kiyotaka Eguchi. 2015. Analysis of User Behavior on Private Chat System. In *2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, Vol. 3. 1–4. https://doi.org/10.1109/WI-IAT.2015.49

[151] A. Upadhyay, A. Chaudhari, and S. Ghale, and S. S. Pawar. 2017. Detection and prevention measures for cyberbullying and online grooming. In *2017 International Conference on Inventive Systems and Control (ICISC)*. 1–4. https://doi.org/10.1109/ICISC.2017.8068605

[152] Joris Van Ouytsel, Narissra M Punyanunt-Carter, Michel Walrave, and Koen Ponnet. 2020. Sexting within young adults' dating and romantic relationships. *Current opinion in psychology* (2020).

[153] Anna Vartapetiance and Lee Gillam. 2014. "Our Little Secret": pinpointing potential predators. *Secur Inform* 3, 1 (Sept. 2014), 3. https://doi.org/10.1186/s13388-014-0003-7

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

[155] Esaú Villatoro-Tello, Antonio Juárez-González, Hugo Jair Escalante, Manuel Montes-y Gómez, and Luis Villaseñor-Pineda. [n.d.]. A Two-step Approach for Effective Detection of Misbehaving Users in Chats. ([n. d.]), 12.

[156] Ashley Marie Walker, Yaxing Yao, Christine Geeng, Roberto Hoyle, and Pamela Wisniewski. 2019. Moving beyond 'one size fits all': research considerations for working with vulnerable populations. *Interactions* 26, 6 (Oct. 2019), 34–39. https://doi.org/10.1145/3358904

[157] Hao Wang, Congxing Cai, Andrew Philpot, Mark Latonero, Eduard H. Hovy, and Donald Metzler. 2012. Data Integration from Open Internet Sources to Combat Sex Trafficking of Minors. In *Proceedings of the 13th Annual International Conference on Digital Government Research (dg.o '12)*. ACM, New York, NY, USA, 246–252. https://doi.org/10.1145/2307729.2307769 event-place: College Park, Maryland, USA.

[158] Terry Winograd. 2006. Shifting viewpoints: Artificial intelligence and human–computer interaction. *Artificial intelligence* 170, 18 (2006), 1256–1258.

[159] Jacob O Wobbrock and Julie A Kientz. 2016. Research contributions in human-computer interaction. *interactions* 23, 3 (2016), 38–44.

[160] Peng Yan, Linjing Li, Weiyun Chen, and Daniel Zeng. 2019. Quantum-Inspired Density Matrix Encoder for Sexual Harassment Personal Stories Classification. In *2019 IEEE International Conference on Intelligence and Security Informatics (ISI)*. 218–220. https://doi.org/10.1109/ISI.2019.8823281 ISSN: null.

[161] 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), e350–e357.

[162] ME Young. 2017. Learning the art of helping: building blocks and techniques. 6. utg.

[163] Patricio Zambrano, Jenny Torres, Luis Tello-Oquendo, Rubén Jácome, Marco E. Benalcázar, Roberto Andrade, and Walter Fuertes. 2019. Technical Mapping of the Grooming Anatomy Using Machine Learning Paradigms: An Information Security Approach. *IEEE Access* 7 (2019), 142129–142146. https://doi.org/10.1109/ACCESS.2019.2942805 Conference Name: IEEE Access.

[164] Aleš Završnik. 2019. Algorithmic justice: Algorithms and big data in criminal justice settings. *European Journal of Criminology* (2019), 1477370819876762.

[165] Jie M Zhang, Mark Harman, Lei Ma, and Yang Liu. 2020. Machine learning testing: Survey, landscapes and horizons. *IEEE Transactions on Software Engineering* (2020).

[166] Yongfeng Zhang and Xu Chen. 2020. Explainable Recommendation: A Survey and New Perspectives. *Foundations and Trends® in Information Retrieval* 14, 1 (2020), 1–101. https://doi.org/10.1561/1500000066 arXiv: 1804.11192.

[167] Z. Zuo, J. Li, P. Anderson, L. Yang, and N. Naik. 2018. Grooming Detection using Fuzzy-Rough Feature Selection and Text Classification. In *2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. 1–8. https://doi.org/10.1109/FUZZ-IEEE.2018.8491591

[168] Ángel Callejas-Rodríguez, Esaú Villatoro-Tello, Ivan Meza, and Gabriela Ramírez-de-la Rosa. 2016. From Dialogue Corpora to Dialogue Systems: Generating a Chatbot with Teenager Personality for Preventing Cyber-Pedophilia. In *Text, Speech, and Dialogue (Lecture Notes in Computer Science)*, Petr Sojka, Aleš Horák, Ivan Kopeček, and Karel Pala (Eds.). Springer International Publishing, 531–539.

A ADDITIONAL TABLES

Table 3. Media Type

Media Types	Counts (Percent)	References
Text	47 (64%)	[9, 11, 23, 24, 26, 27, 30, 36, 37, 39, 40, 42, 43, 46, 49–51, 57, 68, 70, 73, 75, 76, 91, 92, 95, 97–99, 101, 114, 116, 117, 124, 130, 131, 136, 137, 140, 141, 145, 147, 153, 155, 160, 163, 167]
Text and Meta-data	16 (21%)	[7, 23, 29, 35, 41, 54, 64, 69, 79, 84, 85, 88, 90, 115, 118, 123]
Text and Meta-data and image	7 (10%)	[63, 65, 66, 86, 142, 151, 157]
Meta-data	2 (2%)	[150]
Text and image	1 (1%)	[122]

Table 4. Data Types

Datasets	Counts (Percent)	Publications
Perverted Justice	22 (30%)	[11, 23, 27, 30, 41, 51, 70, 79, 84, 90, 95, 97–99, 101, 117, 123, 124, 130, 131, 140, 163]
PAN-2012	16 (22%)	[11, 26, 27, 39, 40, 42, 43, 50, 57, 76, 91, 101, 116, 153, 155, 167]
Combined Chat datasets	10 (14%)	[9, 22–24, 51, 68, 79, 115, 124, 167]
Social Media	8 (11%)	[49, 54, 75, 115, 122, 145, 147, 168]
Advertisements	8 (11%)	[63–66, 86, 141, 142, 157]
Queries	4 (5%)	[46, 85, 114, 137]
Private Chat data	3 (4%)	[37, 69, 150]
SafeCity	3 (4%)	[73, 88, 160]
Games	2 (3%)	[35, 118]
Forums	2 (3%)	[81, 115]
Anonymous Platforms	2 (3%)	[7, 36]
Blogs	2 (3%)	[11, 115]
Generated by Participants	1 (1%)	[136]

Table 5. Ground Truth Annotators

Annotators	Counts (Percent)	Publications
Existing	32 (44%)	[9, 11, 23, 24, 26, 27, 39–43, 51, 57, 70, 73, 76, 86, 91, 97, 99, 101, 115–117, 122, 123, 140, 151, 153, 155, 163, 167, 168]
Outsiders	24 (33%)	Researchers (26%): [7, 30, 36, 37, 49, 50, 54, 65, 66, 75, 79, 88, 95, 124, 130, 131, 140, 142, 147] Moderators (3%): [35, 63] Clinical (1%): [49] Crowd-source (1%): [136]
Auto	17 (23%)	[22, 46, 49, 64, 65, 69, 76, 84, 85, 90, 98, 118, 137, 141, 145, 157, 160]
Insiders	4 (5%)	[49, 75, 81, 145]

Table 6. Features

Features	Counts (Percent)	Publications
Textual	62 (85%)	[7, 9, 11, 22–24, 26, 27, 30, 35–37, 39–41, 43, 49–51, 54, 57, 63, 66, 68, 70, 73, 75, 76, 79, 81, 84, 86, 88, 91, 95, 97–99, 101, 114–118, 122–124, 130, 131, 136, 140–142, 151, 153, 155, 157, 160, 163, 167, 168]
User	20 (27%)	[11, 24, 35, 37, 42, 43, 50, 57, 64–66, 70, 81, 86, 95, 116, 147, 150, 153, 157]
Time/Location	14 (19%)	[41, 63–66, 86, 88, 92, 95, 115, 142, 150, 153, 157]
Semantic	12 (16%)	[7, 22, 30, 40, 41, 90–92, 98, 114, 115, 168]
Style	11 (15%)	[7, 23, 30, 35, 36, 70, 90, 97, 130, 145, 163]
Behavioral	10 (14%)	[23, 35, 51, 57, 79, 95, 116, 130, 150, 153]
Keyword Extraction	10 (14%)	[46, 68, 69, 79, 88, 115, 118, 137, 151, 168]
Syntactic	8 (11%)	[7, 11, 26, 30, 39, 115, 151, 168]
Sentiment	7 (10%)	[23, 24, 30, 35, 90, 145, 151]
Images	5 (7%)	[66, 122, 142, 151, 157]
Network	4 (5%)	[37, 81, 85, 150]
Topic Modeling	3 (4%)	[7, 91, 141, 163]
Relationships	3 (4%)	[23, 85, 95]

Table 7. Approaches

Approaches	Counts (Percent)	Publications
Traditional ML	48 (66%)	Supervised (53%): [7, 9, 11, 23, 24, 26, 27, 30, 35, 37, 42, 43, 51, 54, 57, 70, 79, 84, 91, 95, 97–99, 114–117, 122, 130, 136, 140, 145, 147, 151, 153, 155, 157, 163, 167**?**] Unsupervised (6%): [69, 90, 95, 101, 116, 123, 142, 150] Semi-supervised (3%): [40, 81, 86]
Hand-crafted or Rule-based	13 (18%)	[36, 41, 50, 64, 68, 85, 95, 118, 124, 131, 137, 153, 168]
Deep Learning	11 (15%)	[39, 49, 73, 75, 76, 88, 101, 131, 141, 160, 163]
Graph or Network-based	5 (7%)	[46, 63, 65, 66, 86]
System Architecture	5 (7%)	[85, 92, 118, 142, 157]

Table 8. Algorithms Names

Algorithms	Counts (Percent)	Publications
Support Vector Machine (SVM)	27 (37%)	[7, 24, 26, 27, 30, 35, 37, 40, 43, 51, 54, 81, 91, 98, 114–117, 122, 123, 130, 136, 140, 145, 153, 155, 157]
Naive Bayes (including Multinomial, Gaussian, and Bernoulli) (NB)	16 (22%)	[7, 9, 23, 26, 27, 35, 37, 40, 43, 57, 97–99, 136, 153, 167]
Neural Networks (NN) (including CNN, RNN, LSTM)	15 (21%)	[27, 35, 39, 43, 49, 51, 73, 75, 76, 88, 101, 117, 131, 147, 155, 160, 163]
Regressions (including Logistic (LR), Ridge (RR), Bayesian (BR))	10 (14%)	[7, 9, 27, 27, 35, 43, 70, 114, 124, 167]
K Nearest Neighbor (KNN)	9 (12%)	[35, 43, 51, 81, 95, 98, 117, 142, 145]
Decision Tree (DT)	8 (11%)	[26, 35, 43, 79, 95, 97, 136, 153]
Clustering algorithms (including Agglomerative, AC, KMEANS, AGG, BHC, GMM)	7 (10%)	[30, 69, 79, 81, 86, 123, 150]
Language Model (including BERT)	4 (5%)	[22, 84, 90, 141]
Random Forest (RF)	4 (5%)	[26, 43, 54, 167]
AdaBoost (AB)	3 (4%)	[9, 11, 167]
Contrastive Pessimistic Likelihood Estimation (CPLE)	2 (3%)	[30, 81]
Self-training (ST)	2 (3%)	[145, 151]
Multiple Sequence Alignment (MSA)	1 (1%)	[123]
Local Interpretable Model-Agnostic Explanation (LIME)	1 (1%)	[73]
Linear classifier (LINEAR)	1 (1%)	[163]
RIPPER rule-learning algorithm (RIPPER)	1 (1%)	[95]
Ring Based Classifier (RING)	1 (1%)	[42]
Temporal Relational Semantic Systems (TRSS)	1 (1%)	[41]
Mean Variance	1 (1%)	[50]

Table 9. Granularity Level

Granularity Level	Counts (Percent)	Publications
Users	27 (36%)	[11, 23, 24, 27, 35, 36, 39, 40, 43, 50, 57, 68, 70, 76, 79, 91, 95, 101, 115–118, 123, 130, 150, 153, 155]
Conversations	15 (20%)	[27, 37, 40, 43, 51, 84, 97, 116, 124, 155, 167, 168]
Patterns	9 (12%)	[22, 26, 30, 40–42, 79, 90, 98, 99, 123, 163]
Lines	8 (11%)	[7, 27, 99, 115, 116, 118, 122, 136, 153]
Levels	2 (3%)	[131, 140]

Table 10. Output Types

Output Types	Counts (Percent)	Publications
Binary classification	44 (60%)	[9, 11, 23, 24, 26, 27, 30, 35–37, 39, 42, 43, 49–51, 57, 65, 66, 73, 81, 85, 91, 95, 97, 99, 101, 114–118, 124, 130, 136, 140, 141, 145, 153, 155, 160, 167]
Multi-class classification	24 (33%)	[7, 9, 23, 46, 54, 64, 70, 73, 75, 76, 79, 84, 88, 97, 98, 117, 123, 131, 137, 147, 151, 157, 167, 168]
Clustering	7 (10%)	[66, 68, 69, 79, 86, 142, 150]
Stage detection	5 (7%)	[30, 90, 123, 124, 163]