A Human-Centered Systematic Literature Review of Cyberbullying Detection Algorithms

SEUNGHYUN KIM, Georgia Institute of Technology
AFSANEH RAZI, University of Central Florida, U.S.A
GIANLUCA STRINGHINI, Boston University, U.S.A
PAMELA J. WISNIEWSKI, University of Central Florida, U.S.A
MUNMUN DE CHOUDHURY, Georgia Institute of Technology, U.S.A

Cyberbullying is a growing problem across social media platforms, inflicting short and long-lasting effects on victims. To mitigate this problem, research has looked into building automated systems, powered by machine learning, to detect cyberbullying incidents, or the involved actors like victims and perpetrators. In the past, systematic reviews have examined the approaches within this growing body of work, but with a focus on the computational aspects of the technical innovation, feature engineering, or performance optimization, without centering around the roles, beliefs, desires, or expectations of humans. In this paper, we present a human-centered systematic literature review of the past 10 years of research on automated cyberbullying detection. We analyzed 56 papers based on a three-prong human-centeredness algorithm design framework – spanning theoretical, participatory, and speculative design. We found that the past literature fell short of incorporating human-centeredness across multiple aspects, ranging from defining cyberbullying, establishing the ground truth in data annotation, evaluating the performance of the detection models, to speculating the usage and users of the models, including potential harms and negative consequences. Given the sensitivities of the cyberbullying experience and the deep ramifications cyberbullying incidents bear on the involved actors, we discuss takeaways on how incorporating human-centeredness in future research can aid with developing detection systems that are more practical, useful, and tuned to the diverse needs and contexts of the stakeholders.

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

Additional Key Words and Phrases: cyberbullying detection, human-centered machine learning, literature review, social media

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

Authors' addresses: Seunghyun Kim, skim888@gatech.edu, Georgia Institute of Technology, Atlanta, Georgia; Afsaneh Razi, afsaneh.razi@knights.ucf.edu, University of Central Florida, 4000, Orlando, Florida, U.S.A, 32816; Gianluca Stringhini, Boston University, 02215, Boston, Massachusetts, U.S.A, gian@bu.edu; Pamela J. Wisniewski, University of Central Florida, 4000, Orlando, Florida, U.S.A, pamwis@ucf.edu; Munmun De Choudhury, Georgia Institute of Technology, 30318, Atlanta, Georgia, U.S.A, munmund@gatech.edu.

2573-0142/2021/10-ART325 $15.00
https://doi.org/10.1145/3476066

1 INTRODUCTION

Bullying, a pattern of repeatedly and deliberately harming and humiliating others, specifically those who are smaller, weaker, younger or more vulnerable than the perpetrator, has been a pervasive problem in the society for several decades [172]. With the proliferation of digital social technologies among teens and young adults [146], bullying, once restricted to the school or neighborhood, has now moved into the digital realm. Cyberbullying inflicts unforgettable pain on the victims, with close to two-thirds of U.S. adolescents already having experienced some form of cyberbullying ranging from offensive name-calling to spreading of false rumors [20]. Mental health issues such as anxiety and depression are known to be a result of experiencing bullying as children [172]. The trauma from bullying can also lead to increased suicidal ideation and self-harm [58, 86]. Being bullied at the start of the teenage years has been shown as a potential indicator of the disposition towards borderline personality disorder symptoms [162]. Given its prevalence and long-lasting damage inflicted on young victims of bullying [172], experts agree that cyberbullying is a problem that must be addressed in order to protect the mental health, safety, and well-being of our youth [153].

However, the massive volumes of evolving, real-time, multimodal, heterogeneous, and unstructured social media data makes manual detection of cyberbullying intractable [84]. To address the prevalence of cyberbullying and mitigate the long-term damage caused by these unfortunate events, there has been a growing body of research seeking to develop automated systems to detect cyberbullying incidents. These automated systems aspire and aim to serve a wide range of purposes, ranging from helping prevent the bullying incidents in cyberspace, such as social media, to providing a tool that could support mitigation efforts, such as assisting moderators in online communities to monitor interactions and flag abusive content [34]. In addition the detection mechanisms can also provide support to the victims along with ways to identifying the perpetrators [34].

A few systematic literature reviews in the past have sought to understand the performance and effectiveness of these classifiers from a technical point of view [84, 132, 135]. However, cyberbullying detection is not merely a classification task to identify which and whose content might be abusive towards an individual or group; we posit that *building machine learning models for cyberbullying detection needs to adopt a human-centered perspective*. Reasons range from the sensitivities around the cyberbullying experience [12, 64, 157], its effects on the victim(s) and bystander(s) [42, 159, 172], social stigma [150], impact on the health and functioning of online communities [97], to the potential far-reaching ramifications of cyberbullying incidents on various stakeholders [111].

Recent research in Computer-Supported Cooperative Work and Social Computing (CSCW) has noted that "human-centered paradigms for computing advocate for integrating 'personal, social, and cultural aspects' [77] into the design of technology, and accounting for stakeholders in the creation of technological solutions" [22]. Scholars in the evolving and emergent area of human-centered machine learning have therefore argued that machine learning needs to stay grounded in human needs [22], models need to be built in inclusive ways that adequately represent the diverse experiences of different individuals and minimize biases [122], and that machine learning approaches ought to incorporate interpretability and transparency to not only elucidate its potential for harm [1, 15, 48, 66], but also how data-driven decisions are used in practical scenarios [22, 136]. These practices are important because they provide insights into how machine learning solutions are impacting people, how we should think about existing challenges, and how we should change the way we approach problems so that the models' outcomes align with human and lay interpretations of what said algorithms do and mean [13, 127]. As Amershi et al. [8] rightly noted: "humans are more than "*a source of labels*" and because the process of design should not hinge entirely on the construct of "*the user*" [128], people's involvement with machine learning can take many roles beyond data curation, such as in supporting algorithm selection and tuning, and identifying its points of success

and failure [127]. Articulating these roles and representing them in the development of machine learning algorithms can point to differing agencies between people and the algorithms.

In light of the above, in this paper we argue that building cyberbullying detection models necessitates a deep understanding of a) how cyberbullying can be operationalized based on its theoretical and psychological underpinnings; b) what could be effective ways to represent the varied subjective experiences of cyberbullying within the models given the subjective nature of bullying [135] and how model evaluation needs to look beyond quantifiable metrics to incorporate human feedback, mental models, and social interpretations [13]; and c) who could be the potential stakeholders that could potentially harness the outcomes of such models, and how.

We present a systematic review of the past 10 years of computational research focusing on the development of machine learning models for cyberbullying detection. Adopting the three-prong human-centered algorithm design lens proposed by Baumer [13], in a saturated corpus of 56 papers, we examine how the humans were involved and considered directly or indirectly in the building of these detection algorithms, starting from their design and conceptualization to their evaluation and potential deployment. From a **theoretical standpoint**, we first focus on existing algorithms' alignment with theories of cyberbullying especially in operationalizing acts and incidents of cyberbullying. Then from a **participatory perspective**, we describe if and how existing algorithms have involved the human (or broadly various stakeholders) in data annotation and model evaluation. Finally, from the perspective of **speculative algorithm design**, we shine a light on how researchers have envisioned the usage of existing detection algorithms in real-world scenarios, by who, including consideration of harms and negative consequences. Through this analysis, our review illuminates critical gaps in this research area, that stem from a lack of human-centeredness in algorithm development, and discusses takeaways for future researchers.

2 BACKGROUND

Automated cyberbullying detection is typically a machine learning classification problem where the intent is to classify each abusive or offensive comment, post, message, or image/video as either a bullying or a non-bullying. There have been a few literature reviews in the past to analyze the computational approaches to cyberbullying detection, particularly with a goal to unpack how cyberbullying and its types have been defined from a machine learning perspective [132], what signals in online data serve as the most salient features in classification [84, 132, 135], what types of machine learning methodologies have been adopted [84, 132], how the performances of different models and datasets compare against one another based on standardized metrics like accuracy, precision, and recall [84, 132, 135], and how the paucity of standardized datasets and reliance on manual annotation has hampered reproducibility and replicability [84, 132, 135]. We present these observations about the research area in detail below, followed by how our paper extends these discussions with a human-centered lens.

2.1 Definition and Data-Related Challenges

A major thread within existing review papers has been unpacking the definition of cyberbullying and how to curate a dataset that can detect these incidents with machine learning. Kumar and Sachdeva [84] explored how prior research used various definitions of cyberbullying, ranging from framing and denigration, to outing and impersonation; also see the work of Mahlangu et al. [91] and Nadali et al. [105] on this topic. Other scholars noted that high quality datasets are lacking in this area [105, 135], primarily because of the lack of suitable ground truth data on cyberbullying and therefore a need to rely on manual annotation, which is time-, cost-, and effort-intensive [4, 91]. Vast majority have relied on public social media data [2, 91], which introduces its own biases into the training data because of people's varying self-disclosure behaviors, identity and impression

management goals, and concerns around privacy and context collapse [57]. In fact, Emmery et al. [53] critiqued in their review that there is a reproducibility as well as an evaluation crisis in this research area – most prior work has used small, heterogeneous datasets, without a thorough evaluation of applicability across domains, platforms, and populations. Furthermore, they argued that the positive instances in existing research datasets are often biased to the specific platform of interest, predominantly capturing toxicity, and no other dimensions of bullying (also see [4, 135]).

Importantly, due to the inherently subjective interpretation and experience of cyberbullying incidents [64], researchers have argued that human annotators, used for training data generation, may have different views on which sample is passed as cyberbullying [4]. Subjectivity is not just limited to training data curation; it may exist during the creation of a set of features as well – a fact argued by [4] in their review. This further emphasizes the importance of considering not just the content but also the context of the communication in the datasets, such as history of user activities [38] and the relations among users [29]. While the extent of how much context would affect the performance of such detection models needs further exploration, context has shown to influence how one perceives toxicity [119]. Consequently, Rosa et al. [132], after a systematic review of 22 papers, advocated for establishing well-defined criteria that could help curation of training data and feature engineering, so that the detection models would generalize across datasets, platforms, and contexts. Our paper similarly stresses the need for such harmonious criteria, that we posit can be achieved with a human-centered algorithm design approach.

2.2 Machine Learning Methodology and Evaluation-Related Challenges

A second, complementary set of reviews have focused on the underlying machine learning methodology in cyberbullying detection [2, 83, 105]. A notable survey of prior research by Salawu et al. [135] found the use of many approaches for automated detection, namely, supervised learning, lexicon based, rule based and mixed-initiative approaches. However, many researchers, based on their respective reviews, suggested machine learning methodological improvements, although none considered how these improvements need to stem from real-world scenarios where the algorithms could benefit or potentially harm intended individuals. Kovačević argued that more work needs to be done in terms of taking into account user and contextual aspects of the cyberbullying incidents. Indeed, speaking of context, Lowry et al. [90] emphasized, "*most of these [cyberbullying] studies have glossed over the central issue: the role of ... social media artifacts themselves in promoting cyberbullying.*" Accordingly, Al-Garadi et al. [4] recommended that cyberbullying detection use better feature engineering to capture the rich context of the incidents rather than overly stressing feature selection and machine learning methodological improvements, while Tokunaga [154] suggested careful consideration of user demographic attributes in operationalizing the concept of cyberbullying. However, none of these papers suggested involving the stakeholders of cyberbullying incidents – victims, bystanders, or bullies in capturing this valuable context.

Beyond supervised learning methods – the predominant family of techniques used for cyberbullying detection – researchers have also noted the value of considering other machine learning approaches [2, 4, 103], including unsupervised [44] and semi-supervised techniques [4]. Nevertheless, many researchers also noted that appropriate evaluation needs to go hand in hand with methodological innovation. For instance, most cyberbullying datasets often suffer from significant class-imbalance when the number of positive annotated examples (cyberbullying posts) is much smaller relative to generic social media content [4, 132]. Therefore, researchers have valued careful selection of an evaluation metric that is independent of data skewness, to avoid uncertain results and undesirable outcomes [4]. Suggested evaluation metrics included the F-1 score or the area under receiver-operating characteristic (ROC) curve (AUC), but the existing reviews did not discuss the significance and value of human involvement toward unpacking misclassifications.

Putting it together, these reviews posited that cyberbullying is often inadequately and sometimes misrepresented in the literature with a trickling down effect on training data curation and evaluation of the developed machine learning models. Rosa et al. [132] rightly noted that existing methods, if deployed, are likely to lead to inaccurate systems that would have little real-world application. This paper, that systematically reviews a corpus of 56 papers over the past 10 years that have developed cyberbullying detectors, extends Rosa et al.'s critique. In particular, we consider the human-centered underpinnings of cyberbullying detection algorithms, a hitherto unexplored investigation.

2.3 A Human-Centered Perspective of Machine Learning

Machine learning is increasingly adopted to address societal problems via data-driven decision-making [22], in Fiebrink and Gillies [56]'s words, however, it "often centers on impersonal algorithmic concerns, removed from human considerations such as usability, intuition, effort, and human learning; it is also too often detached from the variety and deep complexity of human contexts in which machine learning may be ultimately applied." Scholars in the CSCW and human-computer interaction (HCI) fields have, therefore, been advocating for a practice that fuses human-centered design with technical work in machine learning systems [13, 101].

Despite the several already existing literature reviews of cyberbullying detection models, we notice that a human-centered analysis has been missing from these literature reviews. First, human-centeredness, in the form of behavioral and social science theories, can provide both prescriptive (helping identify which features might be valuable and why) as well as descriptive knowledge (what do the outcomes of the models mean) in the design of machine learning models [13]. For cyberbullying detection research, these theories can be incredibly valuable [21] – many rich psychological theories like the Control balance theory [32], Dominance theory [140], Just world belief [87], and Crime opportunity theory [32] have been proposed to understand why people engage in cyberbullying as well as elucidate the triadic relationship between victims, perpetrators, and bystanders. These theories can also help to identify the effects of social and technological factors on the participants' thoughts, feelings, and behaviors that can facilitate the development of theoretically-grounded operationalizations of cyberbullying in machine learning models.

Next, Fiebrink and Gillies [56] advocated that examining machine learning from a human-centered perspective includes explicitly recognizing both human work and the human contexts in which machine learning is used. Therefore, considering the subjective experience of cyberbullying [64] and the deeply diverse physical and mental health impacts it leaves on the lives of the victims [71], a human-centered lens can allow us to scrutinize the incorporation of different stakeholder perspectives in the establishment of ground truth in training cyberbullying detection models as well as in evaluating them. A victim could understand an experience totally differently from how the aggressor intended [50]; regardless of whether harm was intended or not. Complementarily, when machine learning models are evaluated by human experts, such as psychologists and mental health professionals in the case of cyberbullying detection, they can help to bridge disconnects between the functionality of the models and their social uses [13].

Third, a human-centered approach to machine learning demands making machine learning more usable and effective for a broader range of stakeholders, including those who would use the outcomes of the machine learning system and those who are affected by them [56]. Many possibilities exist in terms of how cyberbullying detection algorithms may be deployed and used, ranging from prevention to intervention. For instance, Rosa et al. [132] stated that automatic cyberbullying detection can be used to prevent individuals from receiving harmful online content in social networks. At the same time, reflective interfaces can promote users' self-reflection and more pro-social online behaviors, as well as positive online interactions. However, not all errors are created equal – misclassifications may suppress harmless speech, disproportionately stigmatizing

that for particular demographic groups and sometimes even resulting in legal action, whereas in other cases, misclassifications may fail to protect victims subject to actual cyberbullying events or diminish users' trust in the underlying algorithms. A human-centered perspective will allow us to explore these tensions – how algorithms are sensitive to the agency and complexity of the various types of humans using them, and how they might contribute to exacerbating societal biases or lead to unintended negative consequences [151].

To summarize, following Baumer [13], this paper interprets the past literature on cyberbullying detection through the framework of **theoretical**, **participatory**, and **speculative** design [13].

2.4 A Human-Centered Algorithm Design Framework

Baumer [13] conceptualized human-centered algorithm design to engender three key dimensions or strategies – *theoretical, participatory, and speculative design*. These dimensions are neither sequential nor mutually exclusive, but rather, "provide a sense for the range of possibilities" (p. 2, [13]). Therefore, the purpose of this three-prong conceptualization is to ensure that human and social interpretations are incorporated in different ways into the development process of the machine learning algorithm itself. In the sections that follow, we define each of these dimensions:

2.4.1 Theoretical design. According to Baumer, theoretical design incorporates various theories from behavioral and social sciences in the algorithmic design. Scholars have argued that machine learning models are valid only when the theoretical understanding of the concepts under consideration match the operationalization of those same concepts [76]. The theories that are utilized for the design can, therefore, be prescriptive by giving a guideline to *why* certain features should be selected over others for the training of a machine learning model. The use of theories could also could be for descriptive purposes, such as helping the interpretation of the performance of the models. Furthermore, theories in the behavioral and social sciences can help the researcher understand better people's role in the underlying processes operationalized by an algorithm [22], aiding them in their dataset selection, feature selection, and model evaluations.

2.4.2 Participatory design. Unlike theoretical design, participatory design focuses on the involvement of people in the design of the algorithm, as a way to reduce the disconnect between technical solutions and human exposition of the technical solutions. Originating in Scandinavia, this approach has a political dimension of user empowerment and democratization [102]. For others, such as HCI design and usability researchers, it provides a way to involve the stakeholders, designers, researchers, and end-users in the design process to help ensure that the end product meets the needs, desires, and expectations of its intended user base [51]. Therefore, it essentially provides a bridge between people who might be interacting with the development of the system and the ones that designed it. By doing so, in the context of machine learning, this enables an exchange between the possibly varied end users of the algorithm and the designers of the algorithm.

2.4.3 Speculative design. Finally, speculative design relates to provoking important messages, issues, or topics about use of the pertinent algorithm or technology to serve real-world purposes [9]. This design approach therefore helps to identify potential benefits and even unwanted consequences to bridge between the development of the technology and its usage scenarios. It emphasizes that it is important to not just produce artifacts that can be useful, but also be provocative in imagining possible futures with these artifacts. Since it involves going beyond the current problem context to such possible futures, this freedom can facilitate thinking through the ramifications of the algorithm's use in different situations and the (positive or negative) impact on different groups of users or stakeholders.

These three dimensions have shaped the human-centered approach adopted in our literature review, particularly in the generation of the coding rubric that we use to systematically analyze the publications on cyberbullying detection.

3 METHODS

In this section we describe the criteria used to filter relevant publications for our literature review along with how we coded the corpus using the aforementioned human-centeredness framework.

3.1 Corpus Scoping Process

To establish a comprehensive corpus of related literature on cyberbullying detection algorithms, we adopted a method of systematic literature review used by Salawu et al. [135]. We first went through major academic digital libraries – The ACM Digital Library, IEEE Xplore Digital Library, and Springer Link databases for the initial search. These libraries were chosen as one of the key elements of our literature scope was computational approaches; therefore Computer Science publishing focused archival and publication indexing systems were deemed appropriate choices. Combinations of the following keywords were used to identify relevant documents: "cyberbullying," "detection," "detect," and "algorithm." Since we were specifically interested in cyberbullying, we combined cyberbullying and detection/detect/algorithm to ensure that the search results returned publications that studied computational methods to detect cyberbullying.

Once an initial search was completed, we coded each document (paper) for relevancy to our scope based on close reads of the abstract, methodology, and discussion sections. For each relevant document, following standard practice in literature review methodology [113], we then adopted a snowball sampling approach, and went through the reference list to conduct a second pass of search. References that seemed relevant were selected based on the title and abstract. We repeated the process of relevancy coding and going through the references until we reached a stage of saturation, where there were no more new publications being added to the corpus. Since we went through the reference list for each relevant paper, the initial decision of utilizing the three digital libraries does not limit our scope in exploring relevant publications for this literature review, that falls outside of the purview of these three databases.

Prior research has noted such a systematic approach to be very effective at constructing a corpus of publications that are related to the same overall theme [81]. By looking at the reference of the relevant documents at each iteration, we expanded the document base that explores the same domain but possibly published in varied venues (conferences and journals). The robustness of the corpus was also strengthened through this process as we examined the publications that compose the background of each relevant paper relating to cyberbullying detection. In addition, as our iterative approach continued, each iteration resulted in fewer and fewer new relevant documents and more cross-references within the already explored publications, establishing a "closed economy" of pertinent papers that studied computational methods for cyberbullying detection.

Next, each paper was reviewed for inclusion/exclusion using the following criteria:

- The paper needs to either develop or introduce computation methods for cyberbullying detection, using new machine learning techniques or engineering state-of-the-art ones.
- The paper needs to be published between 2010 and 2020. Considering the fast pace of Computer Science research, the time frame was chosen to ensure that the publications were not outdated and largely focused on platforms still in use.
- The paper was only focused on cyberbullying and not other online risk factors, such as hate speech, offensive language, trolling, aggressive, or deviant behaviors, like self-harm.
- The paper needed to use English language data.

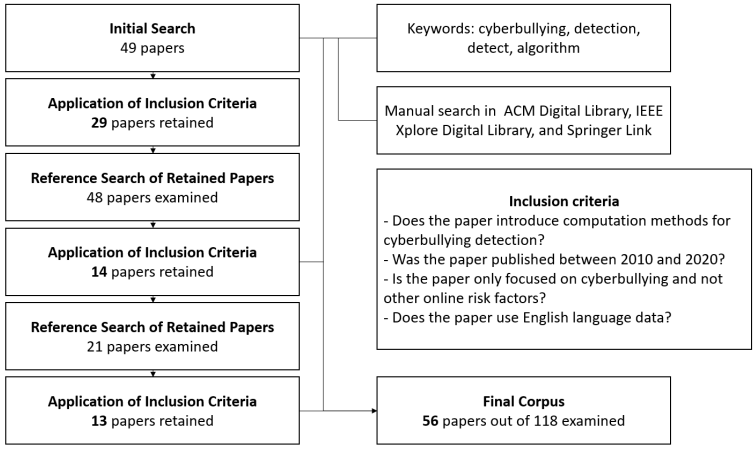

Fig. 1. A general overview of the literature review process. The number of papers examined and retained by search iteration is shown in the boxes, with the number of retained papers and the final number of papers examined in bold.

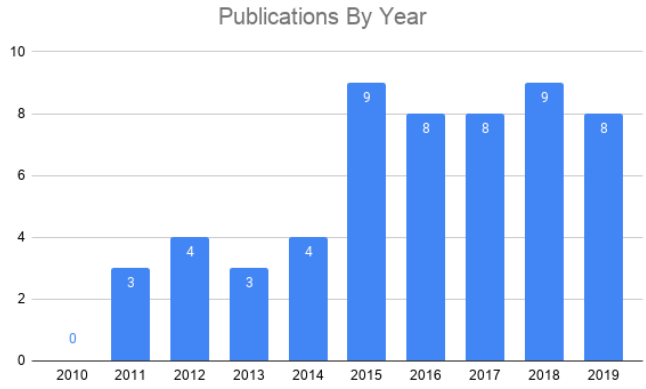

Fig. 2. Count of publications by publication year.

A summary of our entire corpus scoping approach is given in Fig. 1.

3.2 Overview of the Corpus

Using the scoping approach above, the first round of search through the digital libraries yielded a raw set of 49 papers. 20 papers were excluded based on the criteria aforementioned, and an additional 48 papers were examined in the second pass. Out of these 48 papers, 14 were added to the corpus. The third and final pass of the reference lists of the already incorporated papers returned 21 new papers for consideration, of which 13 were relevant. Taken together, **56** papers remained following the application of our inclusion/exclusion criteria. As can be seen in Fig. 2, most of the papers on cyberbullying detection used in this literature review came from more recent past years. When we examined the abstracts of the 56 papers, unsurprisingly, "cyberbullying" was the dominant term, but the abstracts also featured terms like "social media", "bullying", "feature", and "user", giving face validity to the constructed corpus to be reviewed.

Table 1. Coding rubric for the literature review.

Theoretical Design
What were the key aspects of the definition of cyberbullying that the researchers used in their study?
How did the authors decide on what data to use and why?
What were the features used in the development of the model? How were these features chosen?
What machine learning model(s) were used for the task of cyberbullying detection?
Participatory Design
How was the data annotated for training datasets and were humans involved in the process?
How was the model evaluated and did the evaluation consider stakeholder feedback?
Speculative Design
What problems motivated the development of the detection algorithm?
Did the authors speculate potential and unintended future scenarios where their models could be used?
Which stakeholders did the researchers consider in the future use of the developed models?

3.3 Framework and Approach for Coding of the Corpus

Once the saturated corpus of 56 papers was established, we adopted Baumer's three-dimensional conceptual framework on human-centered algorithmic design [13] for qualitatively coding the papers aforementioned in Section 2.4. To align the three dimensions (theoretical, participatory, and speculative design) that capture the human-centeredness of algorithms to a pertinent coding rubric for unpacking the literature on cyberbullying detection, we followed the following approach. We generated questions for each dimension in a deductive way and at the same time used an inductive approach for verification of the validity of the questions. The induction process first selected a random sample of five papers from the corpus to draft a question that would address a shared topic or concept in the subset. Once the question was drafted, another random sample of five papers was examined to see if the established questions were indeed closely related to the papers. Then in the deductive step, we placed the question to the most relevant dimension using their respective definitions (theoretical, participatory, speculative). If no suitable match was found, the inductive and deductive steps were reiterated to reach an alignment.

Table 1 gives a list of the questions that constituted our coding rubric for our literature review. From the table, we note that questions related to incorporating theoretical concepts, models, or relationships within the design of the (cyberbullying detection) algorithm were put under the theoretical design category. When designing a computational model for detecting cyberbullying instances, there are multiple stages in the pipeline where the concept of cyberbullying needs to be operationalized, starting from collecting the data, selecting the features, and deciding on the models to train for the classification task. Therefore, we look at how the definition of cyberbullying was established by examining the use of behavioral or social science theories on bullying and harassment in the studies. We also explore the selection of datasets, features, and models used for the development of the algorithms to further understand *why* they were selected over others.

The participatory design category represents direct involvement of humans in the algorithm design process, the interpretations of the humans in the loop, as well as the end users of the algorithms. The questions were formulated to examine the bridge between the potential end users of the cyberbullying detection algorithm and the researchers. For example, looking into how the dataset for the machine learning models were annotated highlights the importance of incorporating the views and experiences of various actors in cyberbullying in constructing the ground truth for the cyberbullying classifiers – these may include who may have been cyberbullied, those who may have been a bystander in a cyberbullying incident, or psychology/social work experts.

Table 2. Definitions of cyberbullying that were used by the papers in the literature review.

Theoretical Definition	#Articles	References
Aggressiveness/Hostility	82.1%	[5, 17, 23–26, 28, 29, 33, 34, 37, 38, 45, 46, 59, 62, 65, 73, 74, 89, 92, 93, 107, 109, 110, 121, 123–125, 131, 137, 142, 144, 147–149, 152, 155, 156, 158, 164–168, 171]
Repetitiveness	44.6%	[23–25, 28, 33, 34, 37, 38, 46, 65, 73, 74, 89, 107, 109, 123–125, 148, 149, 152, 164, 165, 167, 171]
Imbalance of Power	37.5%	[23–25, 28, 33, 34, 37, 38, 46, 65, 73, 89, 107, 109, 123, 125, 148, 149, 152, 167, 171]

Similarly, interpreting how the performance of these models were evaluated provides insight into how humans interpreted or were (potentially) affected by their success and failure modes.

Lastly, questions related to the envisions and projections that the researchers had about the future use of cyberbullying detection algorithms were put under the speculative design. Here we examine what goals and purposes the researchers aimed to achieve or fulfil, by looking at the main problem statement of the publications and also analyze any potential issues or topics that the studies raised regarding future use and improvement of the detection algorithms. Speculative design also allows us to shine a light on the various stakeholders of the algorithms, when deployed in real-world scenarios, and how those different deployments may impact the stakeholders in varied ways, from the perspective of the benefits-harms calculus.

4 FINDINGS

This section presents the findings of our review of the above corpus of papers. Organized with Baumer [13]'s three-dimensional human-centered algorithm design framework, each of the following subsections interprets the cyberbullying detection algorithms in the papers using the coding rubric described in Table 1.

4.1 Theoretical Design

The first dimension examines how the various theories from behavioral and social sciences were incorporated in the design of the cyberbullying detection algorithms. This category focused on how the theory was used in establishing the definition of cyberbullying as well as the involvement of theory behind the features and dataset.

4.1.1 Definition of Cyberbullying. Several years ago, Dan Olweus [116] introduced the traits of bullying that differentiates it from other types of aggression: repetitive occurrence and clear imbalance between the victim and the aggressor. Many psychology researchers since then have adapted and built upon this definition, such as Raisi and Huang [124]. In our review of the papers, we did find that these three key aspects were heavily and commonly adopted in defining the concept of cyberbullying; see Table 2. Aggressiveness or hostile intent were almost always included in the definition of cyberbullying in 82.1% of the papers with the notion that method of communication to deliver the messages were via internet, cellphones or other devices [142]. However, repetitiveness was not always included in the definition of cyberbullying; only 44.6% of the papers used this attribute in defining cyberbullying. Furthermore, many studies (62.5%) did not include the imbalance of power between the perpetrator and the victim.

Next, there were a few studies that used alternative definitions on top of the concept of cyberbullying. For example, Chatzakou et al.'s study on detecting cyberbullying on Twitter utilized the concept of different roles in the cyberbullying incident such as bully and spammer [23]. Nandhini

and Sheeba's paper defined four different types of cyberbullying: flaming, harassment, racism, and terrorism to establish the framework for their study [149]. The definition of cyberbullying in these cases was constructed in various ways. For example, in some, it was defined concretely through prior literature that focused on defining cyberbullying [34, 144]. Dadvar and De Jong's paper adopted the definition of cyberbullying "as an aggressive, intentional act carried out by a group or individual, using electronic forms of contact (e.g. email and chat rooms) repeatedly or over time against a victim who cannot easily defend herself" (p.121, [34]) from a study on bullying and peer victimization in schools [55]. The study by Singh et al. [144] used the definition from the National Crime Prevention Council which states that cyberbullying is "'when the Internet, cell phones or other devices are used to send or post text or images intended to hurt or embarrass another person'" (p.2, [45]). There were some cases however, where the paper did not specifically or explicitly talk about the definition of cyberbullying used by the authors [16, 27, 36, 40, 106, 108, 129, 139, 168, 169].

4.1.2 Dataset Curation. While the majority of the reviewed studies all used English-based data, there were only a few (7.1%) that explicitly mentioned about and considered the language of the dataset for the ensuing study design and analysis [26, 156, 158, 170]. Further, there were a wide range of datasets used to develop cyberbullying detection models. While a large portion of studies used Twitter (44.6%), some used data from other platforms such as Instagram, YouTube, Slashdot, and MySpace. A few studies (5%) chose particular social media platforms based on sources that listed that top online communities where people experienced cyberbullying [27, 125, 165]. However, there did not seem to be a set or benchmarked dataset that was universally used to drive which platforms were selected as data sources and why; all of the datasets were either from social media platforms (e.g. Twitter, YouTube, Slashdot) or a designated dataset for cybercrimes such as Perverted-Justice [121]. Unlike studies like that of Wijesiriwardene et al. [161], there was little mention of focusing on contextual data when creating the datasets; contextual information such as user activities [38] was later extracted as features. There were also no particular theoretical frameworks from the behavioral or social science literature involved in curating the dataset collected from various platforms. The majority of the studies (71.4%) used the specified platform's API to scrap and collect their own data, thus focusing on primarily public data, essentially indicating an approach to obtain a convenience sample due to ease of access, rather than a sample that is theoretically-justified. While theory cannot always represent real scenarios, the heavy focus on the pragmatism of dataset curation leaves out the consideration of the theories that could further strengthen and improve the curation process.

4.1.3 Feature Selection. Most datasets that are used for cyberbullying detection consist of a great number of possible features to use for training the computational models [132]. For example, one could possibly extract the gender, age, and many other information just from the user profiles [132]. The text that the user posts on the social media platform also has abundant information and many potential ways to represent the data so that it could be utilized in training the classifiers. We looked into how the human decision choices were involved in choosing these features as well as the extent to which existing theories of cyberbullying were incorporated in these decisions.

As the datasets were composed of text sent back and forth between users, textual features were always considered as the mandatory feature for training the models across all publications considered in this review. Textual information ranged from word embeddings to sentiment (23.2%) [17, 23, 24, 28, 65, 93, 108, 139, 156, 167, 168, 168, 169, 171] , part of speech tags (14.3%)[23, 25, 40, 45, 46, 89, 144, 148], to keywords and vocabularies that represented hate, aggression or ill will against another individual (51.8%) [5, 23, 25, 33, 36, 37, 45, 46, 89, 106–110, 124, 125, 129, 137, 142, 144, 152, 155, 156, 158, 164, 165, 168, 168, 169]. The selection of these features was largely data-driven and convenience-driven, rather than theoretically-driven – since these

features can be easily extracted from online data (e.g., using the LIWC program [120]: [27, 28, 142]) or because of the availability of a variety of off-the-shelf language modeling tools (e.g., word embeddings: [17, 28, 93, 139, 167, 168, 168, 169, 171]). That said, the keywords and vocabulary that represented cyberbullying varied; these were often defined by the researchers as a form of domain-knowledge, and therefore a proxy to cyberbullying specific theories. For instance, specific keywords such as 'ugly', 'sick', 'hate', 'fuck' were counted in Yao et al.'s study to detect cyberbullying in Instagram [164], whereas Zhao et al.'s study used a pre-defined set of insulting words based on cosine similarity with word2vec embeddings [169].

Aside from the textual information, metadata about the document such as the number of exclamation marks and the number of words with capital letters were frequently used as features (41.1%) [5, 24, 25, 28, 29, 36–38, 59, 73, 74, 89, 92, 106, 123, 129, 131, 144, 147–149, 152, 156]. While there was a commonly shared characteristic between these studies that they extracted information about the specific dataset entity to further support textual information, there was wide variety in the metadata that the researchers chose, the theoretical or conceptual motivation behind whose selection was rarely clearly articulated in the respective papers. For example, number of negative comments, likes, views were used in a Vine dataset [123] while the number of first/second pronouns, emoticons, ratio of capital letters were extracted in a study using Youtube data [38]; however, *why* each of these metadata meaningfully related to cyberbullying was not articulated.

Similarly, multiple studies (35.7%) incorporated network features and social metadata about the users such as the degree of centrality, tie strength, number of followers, popularity scores to further enrich the feature set [5, 23–25, 29, 33, 34, 36–38, 40, 73, 74, 106, 142, 144, 147, 152, 155, 165]. Number of posts, followers and friends, and days passed since account creation were also used in one study [24]. In another study, gender information was used to train two separate classifiers [35]. Singh et al.'s study to detect multimodal cyberbullying in Instagram used the age/gender of the people in the image as additional features [142] while profile information such as interarrival times were used in Chatzakou et al.'s study [23].

Summarily, there is a large focus on the vocabulary that is used in the posts when it comes to feature selection. Although non-textual features such as user profile data and network features are used, there is a heavy reliance on a bottom-up approach using the content of the posts when it comes to cyberbullying detection, rather than harnessing or unpacking the context of the specific cyberbullying incident. This is an important point to note, given that it is commonly accepted that contextual and temporally varying data is important for cyberbullying detection, as the content alone does not provide enough information for the classifiers [135]. Still, despite the limited use of theory, we do acknowledge that choices of the features in existing research may have been motivated by subjective human observations, which can be valuable to eventually develop or refine cyberbullying theories. In addition, post-hoc analyses of the features that were useful and influential could inspire future work to discover meaningful connections between cyberbullying theories and the features, explaining not only *what* features were important but also *why* they were crucial.

4.1.4 Model Selection. With the sole goal of optimizing for better performance of cyberbullying detection systems, the reviewed studies used a wide range of traditional and state-of-the-art classifiers such as Support Vector Machine (46.4%) [5, 26, 29, 33, 34, 37, 38, 40, 45, 46, 62, 65, 73, 89, 92, 93, 106, 108, 109, 121, 129, 152, 158, 166, 168, 169], Naive Bayes (35.7%) [5, 24, 25, 28, 37, 45, 46, 73, 74, 89, 92, 93, 106, 110, 123, 131, 137, 148, 152, 166], Random Forest (32.1%) [5, 17, 24, 25, 27–29, 38, 59, 89, 93, 106, 123, 152, 164–167], Logistic Regression (23.2%) [17, 26, 28, 29, 40, 89, 92, 93, 106, 123, 152, 164, 171], Tree-based models (10.7%) [23, 24, 27, 37, 123, 147], and AdaBoost (3.6%) [27, 123]. Support Vector Machines (SVM) aim to draw a decision boundary between the classes, maximizing the margin of the separating line; while one of the drawbacks of this approach is that it can be

Table 3. Data annotation methods that were used by the papers in the literature review.

Data Annotation	%Articles	References
External Resources	78.6%	[5, 16, 17, 23–29, 33, 34, 36, 37, 40, 45, 46, 65, 73, 74, 93, 106, 107, 110, 121, 123–125, 129, 131, 142, 144, 147–149, 155, 156, 158, 165–169, 171]
Expert Annotators	12.5%	[5, 27, 29, 124, 125, 158, 166, 171]
Researchers	17.8%	[38, 59, 62, 89, 92, 107–109, 137, 164]

only applied for binary classification, SVMs are often used in cyberbullying detection as it is of the case the study aims for a binary classification of positive or negative bullying [112]. With the assumption that the given features are independent of any given class, Naive Bayes assigns the most likely class when given a feature vector [130]. It has been effective in many fields such as text classification, and medical diagnosis [130]. Random Forest uses a majority-vote out of a combination of tree classifiers to assign the class when given an input vector [117]. Based on a logistic function, the Logistic Regression classifier assigns an estimate between 0 and 1 to a given input; this estimate is therefore used assign the labels to each data entry [82]. Some papers (12.5%) prioritized interpretability in the developed algorithms, and adopted simpler models such as bag of words [45, 156, 168], k-means [27, 131], and k-nearest neighbors [28, 166]. For example, in v Bosque and Garza [156]'s study, a group of lexicon-based approaches were used, based on measurements such as the relative frequency of offensive/swear words, subset of affective words to measure happiness, and lexicon that are grouped into sets cognitive synonyms. These models were compared against statistical, supervised approaches for detecting aggressive text detection on Twitter. However, given the promise of deep learning in recent years, studies have begun to employ models that used neural network of diverse and complex configurations to further improve the detection performance of the models (21.4%) [5, 25, 28, 59, 65, 93, 121, 156, 166–168, 171].

In short, we note a lack of grounding in the cyberbullying literature that could have motivated the selection of these specific models. Scholars have argued that purely optimizing for model performance in machine learning may result in ill-posed problems [7] because such algorithms simply "identify correlations among big data" [7] without fundamentally assessing relationships and inter-dependence between factors, and the mechanisms with which specific attributes may relate to outcomes of interest – insights that are often provided by social science theories. We do note that some of these challenges around robustness, validity, sensitivity, and uncertainty may be addressed with replication and reproducibility studies in the future, and those are encouraged. In addition, model selections are heavily influenced by the problem statement, as they shape the characteristic of the task and the corresponding set of suitable models. Therefore, cyberbullying theories could be used to shape the problem statement and further determine the model selections. Theories may also inspire new machine learning methodologies in cyberbullying detection.

4.2 Participatory Design

In addition to the theoretical incorporation into the design of the models, humans can also participate and get involved directly with the machine learning model pipeline. Obtaining the ground truth for the collected dataset and evaluating the model performance to interpret the results are the two large components in this analysis.

4.2.1 Data Annotation. The insufficient amount of publicly available annotated datasets has always bestowed a challenge to the researchers in the cyberbullying detection field, as they have to establish a way to collect and reasonably label their datasets. Although there were studies where the

researchers themselves labeled the data, thus representing some basic form of human involvement (17.8%) [38, 59, 89, 92, 107–109, 137, 164], other studies relied on external resources to obtain the ground truth for their datasets and often adopted these third-party datasets at face value without further manual introspection of the suitability of the data for the particular task at hand (80.3%). In fact, for these papers, we noted that acquired datasets were often assumed to have high-quality labels, and the acquired annotations were taken to be accurate. Crowdsourcing platforms such as CrowdFlower or Amazon Mechanical Turk were popular as it is known to provide a fast and an easy way to obtain annotations (19.6%) [16, 28, 29, 65, 73, 123, 125, 129, 137, 142, 148]. Given the noisiness of crowd-gathered ground labels [30], most of these papers adopted best practices suggested in crowdsourcing research to assess crowdworkers' task quality and competence [100], but rarely followed them up with additional phases of human verification, triangulation, or a systematic reconciliation of discrepancies when the crowdworkers disagreed. Recruiting college students for labeling was another alternative to outsourcing the annotation, given the overlap in demographics of students and those cyberbullied more frequently (10.7%) [33, 34, 36, 37, 74, 121], but as with crowd-gathered annotations, there was no reflection post-hoc on whether the annotation approach was appropriate for establishing construct validity [114]. Broadly speaking, as can be seen in Table 3, there was a noticeable reliance on some type of external annotators generating the positive and negative cyberbullying examples required for training the cyberbullying detection algorithms.

In contrast to the above, however, there were a handful of studies that involved experts in the annotation process (12.5%) [5, 27, 29, 124, 158, 166, 171]; however, most did not describe the nature of these experts or what attributes qualified them to be an expert at the annotation task. Exceptions include receiving the help of psychology experts in one paper [27] and involving social science expert and behavioral scientists in another [29].

The general set of annotation processes involving humans in some direct or indirect fashion *do* seem to address the challenge of constructing a labeled dataset that can deal with the very sensitive concept of cyberbullying. Still, there remains the question of how researchers can control for the different perceptions of cyberbullying of different people, for instance, the authors of the posts being labeled, the targeted victims, the bystander social media users, community members/moderators, or the platform managers, given the subjectivity of the experience [49] and the diversity of humans who are involved in or impacted by cyberbullying [96]. Importantly, the annotation guidelines in most papers, although often use a certain operationalized or established definition of cyberbullying (96.4%), do not thoroughly account for how the life experiences of the annotators themselves may influence what and how they annotate, because these person-specific experiences are likely to shape how one perceives a given post to be about cyberbullying (or not). The inclusion and reliance on feedback from experts in fields such as psychology or other related social sciences seem to be one way to approach a solution to this issue [27, 171]. However, the absence of adequate and a principled unpacking of these varied perspectives of different stakeholders and the lack of involvement of people with lived experience of cyberbullying indicate a clear shortcoming, since it would only provide one subset of the possible perspectives. Incorporating different perspectives as well as taking advantage of professional guidance from cyberbullying and social science experts will help shape a more comprehensive ground truth in cyberbullying datasets.

4.2.2 Model Evaluation. Moving on to the next part of the machine learning pipeline – model evaluation – conventional machine learning model evaluation metrics such as accuracy [25, 26, 45, 46, 59, 65, 73, 89, 92, 108, 123, 129, 131, 137, 142, 144, 147, 152, 158, 164–168, 168, 171], precision [5, 16, 17, 23–26, 33, 34, 37, 38, 62, 65, 73, 89, 92, 93, 106, 107, 109, 110, 123, 125, 144, 149, 152, 158, 164–169], recall [5, 16, 17, 23–26, 33, 34, 37, 38, 62, 65, 73, 89, 92, 93, 106, 107, 109, 110, 123, 125, 144, 149, 152, 158, 164–169], F1 [5, 25, 27–29, 33, 34, 38, 40, 45, 65, 89, 93, 106, 107, 109, 110, 123, 144,

147–149, 155, 165, 167, 168, 168, 169, 171], and area under receiver-operating characteristic curve (AUC) [5, 17, 23, 25, 26, 28, 36, 37, 40, 59, 142, 158, 164–166] measures were commonly used for evaluating the performance of the cyberbullying detection models. These metrics were computed either on cross-validation data (48.2% of the papers), or held-out test data (37.5%). Some papers combined both, even fewer doing so over multiple experimental runs of their pipeline. While some papers did report multiple performance metrics (91.1%), many relied almost exclusively on one or two metrics such as precision and recall (58.9% and 58.9% papers respectively), without a clear or a rationale, from a human interpretability or understand perspective, why some metrics were prioritized or why certain others were not considered. We rarely found the use of popular metrics from other domains, such as sensitivity, specificity, and positive and negative predictive value.

Importantly, only 3 papers used some of human-evaluation of the models, in fact, just 2 involved experts to assess how well the proposed techniques did in detecting cyberbullying in social media content. But unfortunately, even within these two papers, the background or qualifications of the experts, including how their expertise was defined or assessed were not mentioned [106, 110]. Though not through the involvement of experts, certain other papers provided a qualitative introspection and verification of the results of the detection algorithm, thus constituting an indirect form of human involvement in model performance evaluation. For example, one study provided top covariates and qualitative analyses on the keywords associated with cyberbullying to help interpret the results from their predictive model; the study noted that contrary to literature in public health that states words related to female are positively related to cyberbullying, the results might not be true when controlled for confounders [27]. Another study compared different feature sets to compare and evaluate the top performing feature set for detection [152]. Similarly, Raisi et al.'s study reported the top terms that were censored by the social media platform along with real examples where the annotators labeled the example as non-harassment but the model labeled it as a an example of harassment [125]. Finally, a single study in our reviewed research used a real case study for evaluation – the study conducted a case study at a school located in Spain to evaluate how their model would perform in detecting troll profiles in a real-life scenario [59].

Although not human-centered per se, more nuanced performance measurements such as with true positive [5, 36, 59, 74, 124, 129, 137, 147, 171] or true negative rates [5, 59, 124, 147] have been used in a limited number of papers (21.4%). In addition, about 11% of the papers presented a qualitative error analysis of the classifiers developed [29, 45, 46, 74, 74, 171]. In fact, even for these papers that do provide an error analysis, a human-centered approach to identify potential reasons behind why the algorithm may have misclassified a particular post was absent. For example, Huang et al.'s study does mention how there were some cases of messages including the phrase "stop farting on people" which were not detected as bullying by the textual feature based models, but the authors provided no further detailed error analysis on why the algorithm might have misclassfied [74]. Zhong et al.'s study also provides real examples where their classifier highlighted aggressive comments in a session along with a similarity distribution between the aggressive comment and the caption written by the image poster; they indicated that there was no strong relationship between aggressive comments and the posted content in general [171].

Summarily, the majority of model evaluation approaches lacked detailed interpretation of the performance metrics based on direct human feedback, whether experts or other stakeholders involved in cyberbullying – information that can provide a deeper look at the misclassifications.

4.3 Speculative Design

The third and final dimension of our human-centered review focuses on speculative design considered in this existing body of research. Here we examine what the researchers envisioned about using

the developed cyberbullying detection algorithms, thus providing us with a speculated roadmap of real-world use and deployment in varied real-world scenarios.

4.3.1 Problem Statement and Speculated Use. As noted above, human-centeredness in the form of speculative design considerations evokes critical reflection on the development and role of technology (here, cyberbullying detection algorithms) in the broader societal context [163]. A focus on performance improvements in the cyberbullying detection algorithms, by dint of statistical performance evaluation metrics, has been a central, persistent, and dominant theme in almost all the reviewed papers. Still, different papers adopted slightly different goals in their problem statements, based on their speculated use of the algorithms in the real-world. One of this included developing a machine learning model that has both scalability over vast data size and timeliness in detection [123, 164–166]. For example, Rafiq et al.'s study evaluated the robustness of their model through testing on a very large dataset of 39 million posts [123]. Similarly, Yao's study mentions that 95 million photos and videos are shared on Instagram per day to address how an effective cyberbullying detection system should be able to handle such staggering amount of data [166]. Appearing in about 9% of the papers, timeliness, on the other hand, addressed the problem of establishing a model that could detect a cyberbullying incident in real time, so that the system could raise an alert before the situation deepened or exacerbated.

Other researchers in the reviewed papers considered it paramount for the cyberbullying detection algorithms to adequately process and glean meaningful signals from multimodal data to improve the coverage of the varied types of cyberbullying that may be prevalent online [73, 142, 168, 171]. The rationale was that "'cyberbullying grows bigger and meaner with photos, video' " (p.2091, [142]). And still, several other papers sought to identify different types of key participants involved in cyberbullying incidents, such as the victims or the bullies [24, 25, 59, 108, 109, 137, 147]. Notable is Tahmasbi and Rastegari [152], who noted that a possible approach to cyberbullying detection could identify the cyberbullies instead of cyberbullying messages. Similarly, Nahar et al. [108]'s paper emphasized the need to identify the communication between the predator and the victim. Inferring participant roles on top of detecting cyberbullying was also a distinctive feature in [155].

Central to speculative design is also the emphasis to look beyond technical feasibility, and bring to the fore the assumptions and values embedded in technology, espousing a value fiction approach [52]. While the vast majority of the papers we reviewed adopted a purely technical stance in framing their problem statements and articulating their goals, some did seek to incorporate interdisciplinary knowledge by drawing from fields like psychology [5, 27, 34].

Despite the varied approaches in defining the particular research goals as discussed above, the majority of the reviewed corpus (80.4%) did not speculate how the developed models would be utilized in a real-life scenario, except a handful of exceptions. These exceptions include Tahmasbi and Rastegari's study, wherein the authors speculated how the model could be utilized in an incremental manner to address the vast amount of data when applied to actual social media platforms [152]. There were other papers that envisioned specifically how the models could be used in detecting cyberbullying. The speculations ranged from hypothesizing that governments and governing bodies take action before users become victims of cyberbullying [148, 149], providing support for the victim while tracking the perpetrators [34], giving feedback to stakeholders with authority (parents, law enforcement, etc.) to initiate manual validation of suspected messages [142], inferring trigger comments that cause bullying incidents [171], to offering a way to detect "real" users or those behind fake profiles involved in cyberbullying [59]. Al-garadi et al.'s study, on the other hand mentioned how organization members, non-government organizations as well as crime-prevention foundations could utilize the model, elucidating how different stakeholders could make use of the cyberbullying detection algorithms for their unique needs [5]. That said, it is often the case that

application to real-life scenarios would become an industrial problem and may not necessarily be an expectation for research that develops the detection models. Consequently, considerations of real-life applications are recommended, as researchers are likely to benefit through incorporating speculative design to understand how different stakeholders could benefit from their models.

4.3.2 Speculated Issues in Real-World Use and Deployment of the Detection Models. Next, there were limitations and issues that researchers acknowledged about in the interpretation of the implications of the developed classifiers, which related to their speculated use in the future. 25% of the papers stated that the detection models were built using data from a single specific platform [17, 25, 34, 36, 45, 62, 73, 137, 142, 144, 152, 155, 164, 165]. Therefore, they argued that such domain specificity of the classifiers would limit their application to only those social media platforms that have the same types of features and affordances, limiting generalizable future use. Other papers (16.1%) noted the persistent challenges of curating a high quality ground truth dataset on cyberbullying instances, which would limit training data sizes and therefore hamper practical and robust uses of the algorithms in different real-world scenarios [37, 46, 62, 65, 73, 106, 107, 129, 155]. Further related to data, some papers (5.4%) also acknowledged the limitations of a reliance on purely text-based classification approaches, since words taken at face value can miss nuances in social media language and expression, such as slangs, sarcasm, and irony [26, 38, 158] – an inability to capture these subtleties, the researchers emphasized, can have widespread negative impacts on perpetrators and other stakeholders when the classifiers are deployed in the real-world.

In addition, some of the papers (7.1%) argued that detecting cyberbullying posting alone has little practical value unless the participants involved in the cyberbullying instances are also identified in parallel [107, 108, 152, 155]. Focusing on challenges in real-time use within social media platforms, Sanchez et al.'s study described how a social media platform's existing filtering system could delete posts or comments related to cyberbullying before the researchers could collect the data, drawing attention to the difficulty when a pre-existing cyberbullying detection system eliminates potential positive examples that could be annotated [137].

As noted in Section 2 (Background), the heavy reliance on external annotators for curating training data raises critical questions about how to control for subjectivity of the annotators – a concern noted in previous literature reviews as well [132]. Beyond a notable technical challenge, our review found a small handful of papers speculating that real-world use might be hampered due to the lack of a nuanced approach in the annotation scheme, because the annotators' perspectives simply may not generalize [23, 33, 46, 106, 144, 155]. But even within this small subset, the discussions were limited. Dadvar et al.'s study suggested that the data could be annotated through crowdsourcing as an alternative to capture more varied perspectives [33], while Dinakar et al.'s study showcased semi-supervised learning as a way to overcome the manual annotation of large amounts of data that is likely to be biased by annotators' experiences and views [46]. Tomkins et al.'s study further mentioned a complementary intrinsic challenge with using externally annotations because they are not only subjective but also prone to errors due to a lack of awareness of the situation, even with high inter-rater agreement [155]. Finally, as a significant challenge towards practical use, some papers speculated the implications of misclassifications as well, on the various stakeholders involved in cyberbullying incidents. Notably, Nahar et al.'s study pointed out that false positives and false negatives within cyberbullying detection can have differential impact and interpretation in practical scenarios, and suggested that systems should implement a carefully weighted approach so that while cyberbullying-like posts are not overlooked – incorrectly censoring casual conversations as cyberbullying may hamper online participation and community engagement [106].

Importantly, despite the acknowledgement of these varied issues that would require resolution before real world deployment of the developed cyberbullying detection approaches, we did not

find any paper discuss any potential negative consequences that could arise from implementing the models in social media platforms.

4.3.3 Speculated Users of the Detection Models. As noted before, a variety of stakeholders are involved in any cyberbullying incident, ranging from the victim and bully themselves to the social media platforms. Despite a core focus on optimizing for model performance of the detection models, the reviewed papers identified a wide range of such stakeholders who could potentially use the outcomes of the detection models. Some papers (3.6%) speculated that the developed models could be utilized to support the victims of cyberbullying as a mitigation measure [34, 158]; these papers also recognized a need to track abusive users on social media as a way of cyberbullying prevention. For instance, Dadvar et al.'s study suggests that by following the behaviors of certain users cyberbullying detection systems could identify the victim(s) or perpetrator(s) [34].

Other speculations surrounding stakeholder-centric preventive measures included building tools that provide reflective information to social media users whose language might be construed to be cyberbullying [23, 45], or designing technology supports for community moderators, victim advocacy organizations, and policymakers who seek to prevent cyberbullying behaviors, promote prosocial social engagement online, or fight for and amplify the voices of those victimized by cyberbullying [5, 25, 45, 149]. For instance, researchers speculated that automated detection algorithms can flag messages and help human moderators prioritize a list of potential cyberbullying messages [45]. Additionally, Al-Garadi speculated that the models could be used by various stakeholders, including policy makers and law enforcement bodies [5]. One paper [149] even noted the government to be a potential stakeholder who could benefit from the developed models; these authors discussed that government could take action in social networks using the detecting algorithms, in the form of policy or regulation, to prevent people from becoming victims of cyberbullying.

Perhaps the largest speculated user group in the reviewed papers involved the social media platforms themselves [5, 16, 17, 23–29, 33, 34, 36–38, 40, 45, 46, 59, 62, 65, 73, 74, 89, 92, 93, 106–110, 121, 123–125, 129, 131, 137, 139, 142, 144, 147–149, 152, 155, 156, 158, 164–168, 168, 169, 171]. These papers envisioned that the developed models could be incorporated into content and user filters on the platforms, their recommendation engines, or as standalone affordances that curb cyberbullying from happening altogether. The papers argued that, with the algorithms deployed within the platforms' core functionality, such detection could be more proactive and could be more scalable compared to the state-of-the-art where these judgments are almost always purely manual, and therefore effort-driven and time-consuming. For example, Van Hee et al.'s study mentioned the overload of information on the web and the difficulty of manually monitoring for cyberbullying; the paper points out that the automatic detection of cyberbullying would enhance moderation on social media platforms and help with quicker response to such instances [158]. Dadvar et al.'s study described how the language independence and the adaptiveness of their model could benefit social media platforms as the model could be easily utilized across multiple platforms [36].

That said, these papers rarely adopted a nuanced approach in how exactly a social media platform can harness the potential of the developed algorithms in cyberbullying prevention or use them to curb the negative impacts on its wide range of users, whether the victims, or those who may accurately or mistakenly be flagged as cyberbullies by the underlying algorithm. After all, a platform has little meaning without its users, however, the reviewed papers rarely discussed the impacts these automated services will have on the users beyond the potential business or commercial benefits of timeliness and scalability in detecting cyberbullying.

5 DISCUSSION

In this section, we discuss the implications of our findings of the literature review for each of Baumer [13]'s human-centered design dimensions – theoretical, participatory, and speculative. In our discussions reflecting on the gaps we observed in the prior literature, we also suggest takeaways for future research in cyberbullying detection.

5.1 Theoretical Design: Considerations and Takeaways for Future Research

Theory sits at the crux of social science research; therefore, even for quantitative social scientists, theory is used as a guidance to formulate and test hypotheses [72]. But the algorithmic transformation of theoretical concepts, as is the case for cyberbullying complicates opportunities for theoretical hypothesis testing [13], because the goal of most machine learning models is often to optimize for prediction, instead of generating theoretically-grounded explanations of human behaviors or social phenomena [72], here the cyberbullying experience. That said, theory still has its place in cyberbullying detection research and our literature review noticed several papers where the theory was referenced in operationalizing the concept of cyberbullying. However, found a lack of theoretical engagements, whether in defining the boundaries of cyberbullying, or choosing the dataset, the features, and the machine learning model. In the paragraphs below, we discuss the significance of these missing theoretical engagements, along with considerations for future researchers to close this gap.

5.1.1 Cyberbullying Definition and Operationalization Need to Incorporate Theory. Bullying, including cyberbullying, comes in various forms in multiple environments, and it depends on the experiences of those victimized. Depending on where the bullying happens, however, sometimes the definition of the terminology is modified. Moayed et al., for instance, stated that workplace bullying occurs as result of resolved conflicts [99]. The commonly accepted definition of bullying by Dan Olweus [116] was predominant in our reviewed papers, but it may not be specific enough for the online realm, as argued recently Menesini et al. [94]. That said, the use of theories on bullying to adapt and define the boundaries of cyberbullying is indeed a great starting point as by intuition, cyberbullying is bullying taking place in the setting of the web. In essence, the wide use of Dan Olweus Olweus [116]'s definition of bullying shows how the researchers have been referring pertinent theories for the establishment of the terminology. Such approach will be very beneficial when applying theories from behavioral and social science to developing cyberbullying detection systems. That said, while there has been a shared set of criteria for formally defining the boundaries of cyberbullying in the research we reviewed, we found a lack of consideration in defining hostility or ill-will. Further, researchers often failed to capture the repetitive nature of cyberbullying overtime or the power imbalance within these cyberbullying experiences, and in some cases, they deviated from the theoretical founded definition of bullying altogether. An implication of this departure from theory is that detection algorithms may have a high level of accuracy but not truly be detecting the phenomena of interest, potentially creating false alarms or deploying poorly placed interventions.

To address these gaps, first, in order to account for differences in transferring definition of bullying to specific and unique online contexts from definitions proposed for the offline world, researchers in future work need to identify and understand the affordances of social platforms such as technological affordances as well their social affordances [61, 154]. This will allow an integrated approach to incorporate features like anonymity (technological affordance), with attributes like victims impression and the perpetrators' intention (social affordances) within the detection algorithms. Theories in the behavioral and social sciences can also help the researchers better define different cyberbullying types and its characteristics, for instance, the attributes of repetitiveness and

imbalance of power in the definition of cyberbullying. The rationale behind defining different types of cyberbullying is that each type of bullying manifests differently, and ascribes varied roles to the people involved. For example, there are many circumstances that are all pertinent to cyberbullying such as revenge porn (circulation of sexualized images without the person's consent), trolling (persistent abusive comments), and griefing (harassment in a virtual world or gaming) but each constitutes different types [145]. The definition of cyberbullying, as it sets the framework for the development of the models as well as the data annotation, should consider these subtle differences.

Next, in our review, we found that studies have used an extensive range of features for training their models, exhibiting a comprehensive set of signals that include not only linguistic features but also social network features of the users. Such combination of various types of features may be considered to be the reason driving the commendable performance of several of the existing detection models. However, here as well, we found a lack of theoretical foundations to permeate the engineering and selection of features as well – although a variety of user and post metadata constituted key to cyberbullying detection, appropriate theoretical standards on feature construction and definition was rarely present. Basically, since the availability of features and usefulness of features are subject to the specific traits of the dataset domain, due to epistemological issues around what social media-derived signals really mean when taken out of context [18], and owing to the varied perceptions and nature of cyberbullying across cultural contexts [11], there is a need to establish a set of theoretically-grounded features that can be benchmarked across datasets, cyberbullying types and definitions. Strengthening the theoretical background behind feature selection could also control for potential overfitting from including irrelevant predictors, leading to worse decisions [68]. We suggest that, instead of working in silos in an atheoretical way, computational researchers in cyberbullying detection, therefore, draw upon social science expertise that can provide the theorization needed to close the above noted gap in existing research.

5.1.2 Perspective Matters. The challenge in establishing a commonly accepted universal definition of cyberbullying is closely related to the intrinsic subjectivity; how one perceives cyberbullying is dependent on the role of the individual (bully, victim, bystander) in the cyberbullying incident [64]. The perspective of each victim could also differ; for example, the study by Gualdo et al. [64] who found that those who experienced the traditional offline form of bullying experienced more negative emotional reactions to cyberbullying experiences compared to those that only experienced cyberbullying. Another recent study by Kim et al. [80] showed how differences in perspectives could propagate throughout the development of the classifier for cyberbullying, affecting the ground truth for the dataset and ultimately influencing the performance of the model. None of the papers we reviewed included the elements of subjectivity and perspective differences in the way the data was curated or the machine learning approach developed. Therefore, how can we measure one's intention to hurt someone that can be operationalized into a computational feature in a machine learning model? Will this be from the perspective of the person who says the words or from that of the person who receives the comment? Or would it be from the perspective of a bystander on the platform or the platform's community managers themselves?

Complementarily, in studying bullying, Menesini et al. [94] found that among adolescents across six European countries, young teens had a slightly different perception of power imbalance from the researchers – essentially indicating that the elements of power imbalance between the perpetrator and the victim constitutes not only the inability of the victim to defend themselves from the aggression, but also the agitation caused by the experience, thus rendering power imbalance to be a combination of both the victim's powerlessness and their impression. This begs the question – beyond quantifying some form of aggression as the core signal for training cyberbullying detection models, as has been the case for the reviewed research, how can theoretically grounded foundations

be developed for feature engineering and model selection? We argue that such theorizing can be meaningful not only to examine why a certain feature or model should be considered, but also to support cross-comparison across studies, and importantly post-hoc analysis of when the developed algorithms are likely to be less or more successful in the detection task than others. One could also examine past qualitative studies to augment these computational insights, which have explored further on how one perceives cyberbullying. For example, one qualitative study mentioned how the very definition of cyberbullying differed greatly depending on each individual [95]. Interviews have also revealed that personal experiences seem to often shape the definition of cyberbullying, especially when one has no prior knowledge of different forms of cyberbullying [95]. A systematic review of the qualitative studies on cyberbullying can complementarily provide a detailed analysis about themes and sub-themes that reflect how young people conceptualize cyberbullying [43].

5.1.3 Platform Characteristics Need to be Considered. Using the traditional definition of bullying and adding the medium of such actions as the definition of cyberbullying, as we observed in the reviewed research, while at a glance seems valid, needs to further take into consideration that a different channel of communication also changes the dynamic of how one bullies another. For example, offline bullying could take the form of physical violence or verbal abuse while online bullying is limited to the actions that are possible through online interactions, which varies from platform to platform. Furthermore, each social media platform has their own distinct features [133] which attracts its own unique user segments. Data from a wide range of social media platforms has been used in the reviewed research showcasing generalizability and robustness in detection approaches; however, the diversity across them suggests that the detection models need to account for domain specific traits. Considering the varied ways in which people communicate and talk, in different languages, depending on the social norm of the community that they are part of [39], cyberbullying detection techniques developed in an atheoretical fashion on one dataset may not be effective when evaluated on a dataset from another platform. It should be mentioned however, that past literature have often acknowledged this very aspect of their studies and have stated this as one of their limitations. This allows the readers to consider each study within the specificity of the domain of focus. That said, direct comparison between any two studies, even with a knowledge of their respective limitations may be challenging, given significant demographic differences in terms of *who* uses *which* platform, and structural idiosyncrasies stemming from different platforms' distinct characteristics [133].

Essentially, there needs to be a careful theoretically-justified approach when it comes to setting the boundaries of cyberbullying in a specific online medium, as this lays the foundation for the dataset that is used to train the model to detect cyberbullying.

5.2 Participatory Design: Considerations and Takeaways for Future Research

Next, moving onto the participatory design of human-centered algorithm design, our review indicated that researchers did involve humans in the annotation process to establish ground truth for their datasets. Using detailed instructions for the annotation process, studies have shown strength in resolving or removing data that had major disagreements between the annotators. However, recall that scholars have emphasized the value of involving individuals with lived experiences in the machine learning pipeline, whether to support creating rich and high-quality training data, or to evaluate the outcomes of the models as perceived by potential users of the system [8, 75, 101, 115]. However, our literature review revealed a paucity of meaningful participatory approaches. Instead, we found a singular focus on building the most accurate and well-performing cyberbullying detection model in a way that often meant leaving out the actual and direct participants of the incidents such as the bully and the victim, in the machine learning pipeline, whether in annotation

of the data or the evaluation of the developed models. In the paragraphs below, we discuss the significance of these missing participatory engagements, along with considerations for future researchers to close this gap.

5.2.1 Cultural Differences Need to be Accounted for with Participatory Approaches. There are cultural differences related to perceptions of harassment and aggression among victims as well as bystanders [88]: traditional bullying can take many forms and the prevalence and significance of the behaviors may vary from one cultural setting to another [104]. In European American cultural contexts, for instance, most people are primed with and reinforced for behaving consistent with an independent self-construal (viewing the self as separate from the social context and emphasizing autonomy [141]). However, in Japanese cultural contexts in contrast, most people are primed with and reinforced for behaving in a manner consistent with an interdependent self-construal [141]. Different self-construals may influence a variety of social behaviors, including aggression.

The body of research we reviewed did not explicitly account for cultural background in either annotation of data or evaluation even while involving humans – whether external annotators or experts. Cultural gaps were overlooked even when repurposing an existing corpus – we observed a very large number of papers where a labeled dataset from another paper was used as ground truth at face value without any considerations of who annotated the data and what cultural background they come from. With cultural factors impacting deviant behaviors online as well [10], the manner in which cyberbullying is expressed in social media language needs to be an important aspect in the how the data is annotated and the detection models are evaluated.

5.2.2 Construct Validity Issues Need to be Addressed with Stakeholder Involvement. Next, the reviewed papers employed a variety of external annotators to generate training data for model training, ranging from students and crowdworkers, to social science experts. While such an approach could certainly yield a ground truth dataset that unpacks human subjectivity in interpretations of cyberbullying incidents, there still lies a gap between the perspectives of the victims of the cyberbullying incidents and those of the external annotators. A participatory approach that directly connects with the victims of cyberbullying incidents can help to build a ground truth dataset that justly and accurately reflects people's lived experiences [79], rather than that interpreted by third parties removed from the particular situations.

In fact, the heavy reliance on public social media data in the existing work further underscores the need for this involvement with the victims of cyberbullying. Many social media platforms provide both public and private forms of communication, but as noted by Fiesler et al. [57] public interactions are different in nature compared to private conversations. This is why Aizenkot [3] found their participants to report feeling more cyberbullied through private conversations than through group discourse. However, our review found that little has been explored on the differences between public/private data forms and how they may influence assessments of cyberbullying [3]. When victims and participating actors in cyberbullying are not engaged in ground truth curation, not only does it preclude the inclusion of private conversations in dataset – a consideration important for generalizability – but also misses the context and implicit information that these individuals may provide related to their interpretations of cyberbullying in public and private conversations. Furthermore, according to Chancellor et al. [22], this lack of stakeholder involvement presents ambiguities in delineating both differences between positive and negative examples in machine learning model training, as well as in understanding whether the training data *actually* captures the construct of interest. In fact, Ernala et al. [54] found that, in the context of prediction of mental health states using social media, when ground truth is defined using proxy signals such as hashtag use or self-disclosures without inclusion of self-reported data from individuals with the lived experience of mental illness, construct validity issues lead to significant misclassifications and loss

of credibility in the predictive assessments. Summarily, involvement of the victims of cyberbullying as well as other participating actors such as bystanders is paramount in future studies.

5.2.3 Human Evaluations are Needed along with Computational Measurements. As noted above, a core aspect of participatory design of algorithms centers around involvement of humans in identifying the extent to which the algorithm was successful in achieving what it set out to do [13]. The reviewed set of papers, however, fell short in this front. Critiques of using machine learning for real-world problems [160] acknowledge that performance metrics are an important and necessary aspect of model evaluation because of their ability to provide precise, numerical quantification of performance that is intuitively understandable without needing elaborate or deep understanding of the dataset or the problem domain. They are, however, far from being sufficient – it would be an overstatement to conclude that a highly accurate classifier implies an algorithm that always performs well under any set of circumstances. Scholars have further argued that purely optimizing for model performance in machine learning – as is the case for the reviewed research – may result in ill-posed problems [7] because such algorithms simply "identify correlations among big data" [7] without fundamentally assessing relationships and inter-dependence between factors, and the mechanisms with which specific attributes may relate to outcomes of interest. Different evaluation criteria may additionally encode different value systems [75], but quantitative techniques alone cannot answer the question of which to use.

These issues may be mitigated with involvement of humans, particularly stakeholders of the problem as well as the solution, in evaluation [8]. For instance, although many of the approaches in the papers we reviewed performed well in terms of quantitative performance metrics, their performance from the perspective of their real world users should be evaluated as well. Such human-centered evaluation approaches could involve actual users of the system to test and validate the the model, then the human feedback may be used "in the loop" to tune and improve the model. Next, participatory approaches involving humans could also be used in future work to improve the quality of detection by checking for edge-cases, adding new categories, and so on. Finally, while conducting such evaluations, it will be important to have potential users evaluate the system than researchers, as the latter group might be biased toward the technology that they are building [138].

5.3 Speculative Design: Considerations and Takeaways for Future Research

Finally, algorithms, despite being technical artifacts, do not exist in a vacuum – there is a symbiotic relationship between what an algorithm does and the world it exists in. To this end, as we described above, speculative design provides a lens to look to possible futures and understand the (hidden) impact, influence and future ecosystems that subsume these algorithms [52]. We found the reviewed papers to, however, lack such speculations. Sharing the common problem statement of improving an existing model or implementing a new one altogether, although we did note some diversity of research goals focusing on timeliness, scalability, or multimodality within the detection algorithms, we found them to rarely shine a light on how the models could translate to real-life scenarios, involving diverse stakeholders, and the impacts – positive or negative – the models could have on them. In the paragraphs below, we discuss the significance of these missing speculative engagements, along with considerations for future researchers to close this gap.

5.3.1 A Need to Speculate Who Would Use the Algorithms, Why, and How. Cyberbullying can have long-lasting and varied impact on its victims, as we have noted before [172], and therefore, like other real-world problems [67], misclassifications of models can have varied impacts and bear diverse implications for various stakeholders – whether the victims themselves, the perpetrators, the bystanders or community members, or the social media platforms and moderators. While all errors are equal to a machine learning system, not all errors are equal to all people. Essentially,

human understanding and a human-centered evaluation of model performance that shines a light on the misclassifications, is of paramount importance to make conscious trade-offs between when and for whom to optimize for false positives or for false negatives.

The importance of speculated usage of the models also extend to how the stakeholders could benefit from the classifiers; social media platforms were the dominant stakeholders of past literature, which is not surprising as the researchers in most cases envisioned the models to lead to a real-time detection system. However, there are multiple groups of stakeholders that are directly involved with these online communities, ranging from the users to moderators and administrators. Government officials, policymakers, and law enforcement are also closely related, as they could directly influence the prevention and intervention policies that would affect all social media platforms. In fact, although cyberbullying is not explicitly written in criminal laws, the majority of states in the U.S. have laws that address electronic forms of harassment, providing the responsibility and legal parameters for government and law enforcement involvement in cyberbullying [70]. However, speculative design can enable researchers to think beyond just articulating these different stakeholders. it can empower one to also question what could be specific modes of collaboration with each of them. From the perspective of a potential cyberbullying victim for instance, what should one expect from these automated detection models? On the other hand, how could – or how should – moderators of social media platforms use these models when identifying cyberbullying incidents and cyberbullies? What would be acceptable interventions and who decides what is acceptable?

5.3.2 A Need to Weigh on Negative Consequences and the Ethics of Detection. Finally, the noticeable lack of speculations on how the models would be used in real-life scenarios illustrates how past literature fell short in illuminating potential benefits and harms to different stakeholders. It is easy for one to assume that automated machine learning decisions are omnipotent – however, a consideration of negative consequences is critical in cyberbullying detection given the deep implications for the victims, perpetrators, and bystanders [19]. Overlooking negative impacts could result in considering only the positive side of the models, and could lead to damaging negative impacts [69]. For example, a user might be wrongfully flagged as a cyberbully by a detection model. If this model is implemented in the real world, the consequences could be far-reaching. Depending on the intervention and content moderation policy of the platform, this wrongfully flagged user, for instance, could have their posts sanctioned, or worse, be permanently banned, with no more access to the services of the platform. In fact, if banning is aggressively implemented with high rates of false positives, it can not only be stigmatizing, but also can lead to users either self-censoring their speech or leaving the social platform altogether [134]. Similar negative consequences may be envisioned for the victims of the cyberbullying incidents as well. A false negative in this case could potentially result in overlooking a victim of cyberbullying, missing the opportunity for moderators to intervene or support the individual who might be under distress and difficult circumstances. Speculating such negative consequences in future work can help adopters of the machine learning models to foresee these intricacies and implications of implementation rather than blindly incorporating the models in practical applications.

In addition, researchers need to speculate potential biases of their developed models – a discussion that was missing in the reviewed papers. Potential bias may originate in the dataset – whether its source, filtering and curation strategy used by the researchers, or the manner in which ground truth data is annotated [75]. In fact, recent studies have revealed that existing abusive speech detection systems carry significant amount of unintended bias against certain groups of users, demographics, languages, dialects, terms, phrases, or topics [41, 47, 118]. For instance, terms like "gay" and "jew," present in an informative or conversational context, tend to incline the model predictions towards "hate speech" or "high toxicity." Similarly, Twitter posts in African-American Vernacular English

have been shown to be more likely to be classified as abusive or offensive compared to other postings [41]. As such groups are often minoritized with respect to the general online population, attempts to maximize the detection accuracy and evaluating approaches solely on conventional model performance metrics may exacerbate the biases even further or give rise to newer, previously unconceived discriminatory traits. While such findings could be insightful in providing warning indicators to the stakeholders of cyberbullying, it still remains critical that researchers acknowledge any potential bias presented through developing the models. For instance, researchers can consider approaches like "data statements" [14] and "model cards" [98] that have been advocated recently to being transparency to ML systems. Together, data statements and models cards not only can provide benchmarked evaluation of a cyberbullying detection system in a variety of conditions, such as across different cultural, demographic, phenotypic, or intersectional groups, but also help to disclose the context in which the data and the models are intended to be used, the details of the performance evaluation procedures, how the system might be appropriately deployed, what biases might be reflected in practical uses of the system, as well as what harms may be perpetuated.

Complementarily, these efforts could be augmented by research that have explored techniques to mitigate potential unwanted biases in the models such as by adding a fairness constraint [60], odd post-processing technique [143], incorporating of a fairness score when optimizing for accuracy [6], and penalizing for unfairness in the training of the models [126]. Specifically, studies have attempted to achieve algorithmic fairness during the pre-processing of the data, training of the algorithms, and processing of results [143]. The focus of achieving fairness aims to reduce the amount of variation in the performance of the algorithm due to certain attributes such as race or gender [6]. We advocate future research to not only speculate sources of biases, but also to correct the machine learning approaches so that they could be used in inclusive way in the real-world.

Last but not the least, we discuss how speculative design may also build ethical practices into the development of the machine learning pipeline for cyberbullying detection. Although throughout this paper, we argued the need for directly involving participating actors in cyberbullying incidents – such as victims, perpetrators, and bystanders – we suggest caution in doing so. Social science research has indicated that social desirability bias remains a significant issue in collecting data on the behavioral patterns of perpetrators and bystanders: respondents who engage in socially undesirable acts tend to under-report their participation, whereas those who engage in socially desirable acts tend to over-report their participation, in order to be viewed favorably respectively [63]. Second, research on victimization is constrained by the ethical need to avoid "harming" the participants [31]. Therefore, by adopting speculative design, researchers can focus on adopting study designs that still preserve human-centeredness without serving as a potential cause or source of harm.

Nevertheless, we do recognize the potential risk that speculative design approaches may post to individuals with the lived experience of cyberbullying, such as perpetually casting certain identity groups as "victims", such as people of color, LGBTQ+ individuals, and so on. In using speculative design for cyberbulling detection, we therefore suggest guidelines suggested by Jo and Gebru [78] who emphasize considering issues such as consent, power, inclusivity, transparency, and ethics and privacy in the data curation practices and approaches to develop machine learning pipelines. For instance, in this work, Jo and Gebru suggest democratizing the data collection process, and giving agency to minority groups to represent themselves. At the same time, guidelines from the psychology field may be adopted as part of the speculative design exercises that ensure that when people who have been cyberbullied are involved in research, their identities are adequately protected and support resources are deployed as protective measures [85].

6 CONCLUSION

In this paper we conducted a systematic review of the past literature on automated cyberbullying detection models. After establishing a corpus of relevant documents to cyberbullying detection, we analyzed the human involvement in the development of these models using an established human-centered algorithmic design framework [13]. Specifically, we reviewed the past research in terms of their considerations for theoretical, participatory, and speculative design. Our review revealed that despite extensive research on developing cyberbullying detection models that optimize for statistical performance and methodological innovation, there were clear gaps in terms of a) how the complex phenomenon of cyberbullying was defined and operationalized from a theoretical-grounding perspective; b) how a lack of involvement of stakeholders of bullying in data curation exposed potential for construct validity issues; and c) how poor speculation of the uses and users of the algorithms not only hampered model evaluation in real-world scenarios, but opened up opportunities for harm to various participating actors of cyberbullying. We concluded with guidelines on how a human-centered approach can help to address these pervasive concerns in this important research area within social computing.

7 ACKNOWLEDGEMENTS

This study is supported by the United States National Science Foundation under grant IIP-1827700. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] Rediet Abebe and Kira Goldner. 2018. Mechanism design for social good. *AI Matters* 4, 3 (2018), 27–34.
[2] S Aboud El-Seoud, Nadine Farag, and Gerard McKee. 2020. A review on non-supervised approaches for cyberbullying detection. *International Journal of Engineering Pedagogy* 10, 4 (2020), 25–34.
[3] Dana Aizenkot. 2020. Cyberbullying experiences in classmates âĂŸWhatsApp discourse, across public and private contexts. *Children and Youth Services Review* 110 (2020), 104814.
[4] Mohammed Ali Al-Garadi, Mohammad Rashid Hussain, Nawsher Khan, Ghulam Murtaza, Henry Friday Nweke, Ihsan Ali, Ghulam Mujtaba, Haruna Chiroma, Hasan Ali Khattak, and Abdullah Gani. 2019. Predicting cyberbullying on social media in the big data era using machine learning algorithms: review of literature and open challenges. *IEEE Access* 7 (2019), 70701–70718.
[5] Mohammed Ali Al-Garadi, Kasturi Dewi Varathan, and Sri Devi Ravana. 2016. Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network. *Computers in Human Behavior* 63 (10 2016), 433–443. https://doi.org/10.1016/j.chb.2016.05.051
[6] Jamal Alasadi, Ramanathan Arunachalam, Pradeep K Atrey, and Vivek K Singh. 2020. A Fairness-Aware Fusion Framework for Multimodal Cyberbullying Detection. In *2020 IEEE Sixth International Conference on Multimedia Big Data (BigMM)*. IEEE, 166–173.
[7] Mark Alber, Adrian Buganza Tepole, William R Cannon, Suvranu De, Salvador Dura-Bernal, Krishna Garikipati, George Karniadakis, William W Lytton, Paris Perdikaris, Linda Petzold, et al. 2019. Integrating machine learning and multiscale modelingâĂŤperspectives, challenges, and opportunities in the biological, biomedical, and behavioral sciences. *NPJ digital medicine* 2, 1 (2019), 1–11.
[8] Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. Power to the people: The role of humans in interactive machine learning. *Ai Magazine* 35, 4 (2014), 105–120.
[9] James Auger. 2013. Speculative design: crafting the speculation. *Digital Creativity* 24, 1 (2013), 11–35.
[10] Jieun Baek and Lyndal M Bullock. 2014. Cyberbullying: a cross-cultural perspective. *Emotional and behavioural difficulties* 19, 2 (2014), 226–238.
[11] Christopher P Barlett, Douglas A Gentile, Craig A Anderson, Kanae Suzuki, Akira Sakamoto, Ayuchi Yamaoka, and Rui Katsura. 2014. Cross-cultural differences in cyberbullying behavior: A short-term longitudinal study. *Journal of cross-cultural psychology* 45, 2 (2014), 300–313.
[12] Julia Barlińska, Anna Szuster, and Mikołaj Winiewski. 2013. Cyberbullying among adolescent bystanders: Role of the communication medium, form of violence, and empathy. *Journal of Community & Applied Social Psychology* 23, 1 (2013), 37–51.

[13] Eric PS Baumer. 2017. Toward human-centered algorithm design. *Big Data & Society* 4, 2 (2017), 2053951717718854.
[14] 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.
[15] Ruha Benjamin. 2019. Assessing risk, automating racism. *Science* 366, 6464 (2019), 421–422.
[16] Jacob L. Bigelow, April Edwards, and Lynne Edwards. 2016. Detecting cyberbullying using latent semantic indexing. In *Proceedings of the 1st International Workshop on Computational Methods for CyberSafety, CyberSafety 2016*. Association for Computing Machinery, Inc, 11–14. https://doi.org/10.1145/3002137.3002144
[17] Tazeek Bin Abdur Rakib and Lay Ki Soon. 2018. Using the Reddit Corpus for Cyberbully Detection. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 10751 LNAI. Springer Verlag, 180–189. https://doi.org/10.1007/978-3-319-75417-8{_}17
[18] Danah Boyd and Kate Crawford. 2012. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society* 15, 5 (2012), 662–679.
[19] Wanda Cassidy, Chantal Faucher, and Margaret Jackson. 2013. Cyberbullying among youth: A comprehensive review of current international research and its implications and application to policy and practice. *School psychology international* 34, 6 (2013), 575–612.
[20] Pew Research Center. 2018. A Majority of Teens Have Experienced Some Form of Cyberbullying. https://www.pewresearch.org/internet/2018/09/27/a-majority-of-teens-have-experienced-some-form-of-cyberbullying/
[21] Tommy KH Chan, Christy MK Cheung, and Zach WY Lee. 2020. Cyberbullying on Social Networking Sites: A Literature Review and Future Research Directions. *Information & Management* (2020), 103411.
[22] Stevie Chancellor, Eric PS Baumer, and Munmun De Choudhury. 2019. Who is the "Human" in Human-Centered Machine Learning: The Case of Predicting Mental Health from Social Media. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–32.
[23] Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. 2017. Mean birds: Detecting aggression and bullying on Twitter. In *WebSci 2017 - Proceedings of the 2017 ACM Web Science Conference*. Association for Computing Machinery, Inc, New York, New York, USA, 13–22. https://doi.org/10.1145/3091478.3091487
[24] Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. 2019. Detecting aggressors and bullies on twitter. In *26th International World Wide Web Conference 2017, WWW 2017 Companion*. International World Wide Web Conferences Steering Committee, New York, New York, USA, 767–768. https://doi.org/10.1145/3041021.3054211
[25] Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, Athena Vakali, and Nicolas Kourtellis. 2019. Detecting Cyberbullying and Cyberaggression in Social Media. *ACM Transactions on the Web* 13, 3 (10 2019), 1–51. https://doi.org/10.1145/3343484
[26] Vikas S. Chavan and S. S. Shylaja. 2015. Machine learning approach for detection of cyber-aggressive comments by peers on social media network. In *2015 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2015*. Institute of Electrical and Electronics Engineers Inc., 2354–2358. https://doi.org/10.1109/ICACCI.2015.7275970
[27] Lu Cheng, Ruocheng Guo, and Huan Liu. 2019. Robust cyberbullying detection with causal interpretation. In *The Web Conference 2019 - Companion of the World Wide Web Conference, WWW 2019*. Association for Computing Machinery, Inc, 169–175. https://doi.org/10.1145/3308560.3316503
[28] Lu Cheng, Ruocheng Guo, Yasin Silva, Deborah Hall, and Huan Liu. 2019. Hierarchical attention networks for cyberbullying detection on the instagram social network. In *SIAM International Conference on Data Mining, SDM 2019*. Society for Industrial and Applied Mathematics Publications, 235–243. https://doi.org/10.1137/1.9781611975673.27
[29] Lu Cheng, Jundong Li, Yasin N. Silva, Deborah L. Hall, and Huan Liu. 2019. XBully: Cyberbullying detection within a multi-modal context. In *WSDM 2019 - Proceedings of the 12th ACM International Conference on Web Search and Data Mining*. Association for Computing Machinery, Inc, 339–347. https://doi.org/10.1145/3289600.3291037
[30] Anand Inasu Chittilappilly, Lei Chen, and Sihem Amer-Yahia. 2016. A survey of general-purpose crowdsourcing techniques. *IEEE Transactions on Knowledge and Data Engineering* 28, 9 (2016), 2246–2266.
[31] James J Clark and Robert Walker. 2011. Research ethics in victimization studies: Widening the lens. *Violence against women* 17, 12 (2011), 1489–1508.
[32] Lawrence E Cohen and Marcus Felson. 1979. Social change and crime rate trends: A routine activity approach. *American sociological review* (1979), 588–608.
[33] Maral Dadvar, Franciska De, Jong Roeland, and Ordelman Dolf Trieschnigg. 2012. *Improved Cyberbullying Detection Using Gender Information*. Technical Report. http://www.noswearing.com/dictionary
[34] Maral Dadvar and Franciska De Jong. 2012. Cyberbullying detection: A step toward a safer internet yard. In *WWW'12 - Proceedings of the 21st Annual Conference on World Wide Web Companion*. ACM Press, New York, New York, USA, 121–125. https://doi.org/10.1145/2187980.2187995

[35] Maral Dadvar, FMG de Jong, Roeland Ordelman, and Dolf Trieschnigg. 2012. Improved cyberbullying detection using gender information. In *Proceedings of the Twelfth Dutch-Belgian Information Retrieval Workshop (DIR 2012)*. University of Ghent.

[36] Maral Dadvar, Dolf Trieschnigg, and Franciska De Jong. 2013. *Expert knowledge for automatic detection of bullies in social networks*. Technical Report. 57–64 pages.

[37] Maral Dadvar, Dolf Trieschnigg, and Franciska De Jong. 2014. Experts and machines against bullies: A hybrid approach to detect cyberbullies. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 8436 LNAI. Springer Verlag, 275–281. https://doi.org/10.1007/978-3-319-06483-3{_}25

[38] Maral Dadvar, Dolf Trieschnigg, Roeland Ordelman, and Franciska De Jong. 2013. Improving cyberbullying detection with user context. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 7814 LNCS. Springer, Berlin, Heidelberg, 693–696. https://doi.org/10.1007/978-3-642-36973-5_62

[39] Cristian Danescu-Niculescu-Mizil, Robert West, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. No country for old members: User lifecycle and linguistic change in online communities. In *Proceedings of the 22nd international conference on World Wide Web*. 307–318.

[40] Harsh Dani, Jundong Li, and Huan Liu. 2017. Sentiment Informed Cyberbullying Detection in Social Media. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 10534 LNAI. Springer Verlag, 52–67. https://doi.org/10.1007/978-3-319-71249-9_4

[41] Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial Bias in Hate Speech and Abusive Language Detection Datasets. In *Proceedings of the Third Workshop on Abusive Language Online*. Association for Computational Linguistics, Florence, Italy, 25–35. https://doi.org/10.18653/v1/W19-3504

[42] Katie Davis, David P Randall, Anthony Ambrose, and Mania Orand. 2015. 'I was bullied too': stories of bullying and coping in an online community. *Information, Communication & Society* 18, 4 (2015), 357–375.

[43] Rebecca Dennehy, Sarah Meaney, Kieran A Walsh, Carol Sinnott, Mary Cronin, and Ella Arensman. 2020. Young people's conceptualizations of the nature of cyberbullying: A systematic review and synthesis of qualitative research. *Aggression and violent behavior* 51 (2020), 101379.

[44] Michele Di Capua, Emanuel Di Nardo, and Alfredo Petrosino. 2016. Unsupervised cyber bullying detection in social networks. In *2016 23rd International conference on pattern recognition (ICPR)*. IEEE, 432–437.

[45] Karthik Dinakar, Birago Jones, Catherine Havasi, Henry Lieberman, and Rosalind Picard. 2012. Common Sense Reasoning for Detection, Prevention, and Mitigation of Cyberbullying. *ACM Transactions on Interactive Intelligent Systems* 2, 3 (9 2012), 1–30. https://doi.org/10.1145/2362394.2362400

[46] Karthik Dinakar, Roi Reichart, and Henry Lieberman. 2011. Modeling the detection of textual cyberbullying. In *AAAI Workshop - Technical Report*, Vol. WS-11-02. 11–17. www.aaai.org

[47] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 67–73.

[48] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).

[49] Rebecca Dredge, John Gleeson, and Xochitl De la Piedad Garcia. 2014. Cyberbullying in social networking sites: An adolescent victimâĂŹs perspective. *Computers in human behavior* 36 (2014), 13–20.

[50] Rebecca Dredge, John Gleeson, and Xochitl De La Piedad Garcia. 2014. Presentation on Facebook and risk of cyberbullying victimisation. *Computers in Human Behavior* 40 (8 2014), 16–22. https://doi.org/10.1016/j.chb.2014.07.035

[51] Ana Maria Bustamante Duarte, Nina Brendel, Auriol Degbelo, and Christian Kray. 2018. Participatory design and participatory research: An HCI case study with young forced migrants. *ACM Transactions on Computer-Human Interaction (TOCHI)* 25, 1 (2018), 1–39.

[52] Anthony Dunne and Fiona Raby. 2013. *Speculative everything: design, fiction, and social dreaming*. MIT press.

[53] Chris Emmery, Ben Verhoeven, Guy De Pauw, Gilles Jacobs, Cynthia Van Hee, Els Lefever, Bart Desmet, Véronique Hoste, and Walter Daelemans. 2020. Current limitations in cyberbullying detection: On evaluation criteria, reproducibility, and data scarcity. *Language Resources and Evaluation* (2020), 1–37.

[54] Sindhu Kiranmai Ernala, Michael L Birnbaum, Kristin A Candan, Asra F Rizvi, William A Sterling, John M Kane, and Munmun De Choudhury. 2019. Methodological gaps in predicting mental health states from social media: Triangulating diagnostic signals. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–16.

[55] Dorothy L Espelage and Susan M Swearer. 2003. Research on school bullying and victimization: What have we learned and where do we go from here? *School psychology review* 32, 3 (2003), 365–383.

[56] Rebecca Fiebrink and Marco Gillies. 2018. Introduction to the special issue on human-centered machine learning.
[57] Casey Fiesler, Michaelanne Dye, Jessica L Feuston, Chaya Hiruncharoenvate, Clayton J Hutto, Shannon Morrison, Parisa Khanipour Roshan, Umashanthi Pavalanathan, Amy S Bruckman, Munmun De Choudhury, et al. 2017. What (or who) is public? Privacy settings and social media content sharing. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 567–580.
[58] Helen L Fisher, Terrie E Moffitt, Renate M Houts, Daniel W Belsky, Louise Arseneault, and Avshalom Caspi. 2012. Bullying victimisation and risk of self harm in early adolescence: longitudinal cohort study. *bmj* 344 (2012), e2683.
[59] Patxi Galán-García, José Gaviria de la Puerta, Carlos Laorden Gómez, Igor Santos, and Pablo García Bringas. 2016. Supervised machine learning for the detection of troll profiles in twitter social network: Application to a real case of cyberbullying. *Logic Journal of the IGPL* 24, 1 (2016), 42–53.
[60] Oguzhan Gencoglu. 2020. Cyberbullying Detection with Fairness Constraints. *arXiv preprint arXiv:2005.06625* (2020).
[61] James J Gibson. 1977. The theory of affordances. *Hilldale, USA* 1, 2 (1977).
[62] Kim D. Gorro, Mary Jane G. Sabellano, Ken Gorro, Christian Maderazo, and Kris Capao. 2018. Classification of Cyberbullying in Facebook Using Selenium and SVM. In *2018 3rd International Conference on Computer and Communication Systems, ICCCS 2018*. Institute of Electrical and Electronics Engineers Inc., 233–238. https://doi.org/10.1109/CCOMS.2018.8463326
[63] Pamela Grimm. 2010. Social desirability bias. *Wiley international encyclopedia of marketing* (2010).
[64] Ana M Giménez Gualdo, Simon C Hunter, Kevin Durkin, Pilar Arnaiz, and Javier J Maquilón. 2015. The emotional impact of cyberbullying: Differences in perceptions and experiences as a function of role. *Computers & Education* 82 (2015), 228–235.
[65] John Hani, Mohamed Nashaat, Mostafa Ahmed, Zeyad Emad, Eslam Amer, and Ammar Mohammed. 2019. Social media cyberbullying detection using machine learning. *International Journal of Advanced Computer Science and Applications* 10, 5 (2019), 703–707. https://doi.org/10.14569/ijacsa.2019.0100587
[66] Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. 2020. Towards a critical race methodology in algorithmic fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 501–512.
[67] Morten Steen Hansen, Per Fink, Morten Frydenberg, Marie-Louise Oxhøj, Lene Søndergaard, and Povl Munk-Jørgensen. 2001. Mental disorders among internal medical inpatients: prevalence, detection, and treatment status. *Journal of Psychosomatic Research* 50, 4 (2001), 199–204.
[68] Douglas M Hawkins. 2004. The problem of overfitting. *Journal of chemical information and computer sciences* 44, 1 (2004), 1–12.
[69] Brent Hecht, Lauren Wilcox, Jeffrey P Bigham, Johannes Schöning, Ehsan Hoque, Jason Ernst, Yonatan Bisk, Lana Yarosh, Bushra Amjam, and Cathy Wu. 2018. It's time to do something: Mitigating the negative impacts of computing through a change to the peer review process. *ACM Future of Computing Blog* (2018).
[70] Sameer Hinduja and Justin W Patchin. 2012. School law enforcement and cyberbullying. *Cyberbullying prevention and response: Expert perspectives* (2012), 161–184.
[71] Dianne L Hoff and Sidney N Mitchell. 2009. Cyberbullying: Causes, effects, and remedies. *Journal of Educational Administration* (2009).
[72] Jake M Hofman, Amit Sharma, and Duncan J Watts. 2017. Prediction and explanation in social systems. *Science* 355, 6324 (2017), 486–488.
[73] Homa Hosseinmardi, Sabrina Arredondo Mattson, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra. 2015. Detection of Cyberbullying Incidents on the Instagram Social Network. *MobiSys* (3 2015), 2014. http://arxiv.org/abs/1503.03909
[74] Qianjia Huang, Vivek Kumar Singh, and Pradeep Kumar Atrey. 2014. Cyber Bullying Detection Using Social and Textual Analysis. In *Proceedings of the 3rd International Workshop on Socially-Aware Multimedia - SAM '14*. ACM Press, New York, New York, USA, 3–6. https://doi.org/10.1145/2661126.2661133
[75] Ben Hutchinson and Margaret Mitchell. 2019. 50 years of test (un) fairness: Lessons for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 49–58.
[76] Abigail Z Jacobs and Hanna Wallach. 2019. Measurement and fairness. *arXiv preprint arXiv:1912.05511* (2019).
[77] 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.
[78] 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.
[79] David Jurgens, Libby Hemphill, and Eshwar Chandrasekharan. 2019. A Just and Comprehensive Strategy for Using NLP to Address Online Abuse. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 3658–3666. https://doi.org/10.18653/v1/P19-1357
[80] Seunghyun Kim, Afsaneh Razi, Gianluca Stringhini, Pamela Wisniewski, and Munmun De Choudhury. forthcoming. 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.
[81] Barbara Kitchenham and Stuart Charters. 2007. Guidelines for performing systematic literature reviews in software engineering. (2007).
[82] David G Kleinbaum, K Dietz, M Gail, Mitchel Klein, and Mitchell Klein. 2002. *Logistic regression*. Springer.
[83] Ana Kovačević. 2014. Cyberbullying detection using web content mining. In *2014 22nd Telecommunications Forum Telfor (TELFOR)*. IEEE, 939–942.
[84] Akshi Kumar and Nitin Sachdeva. 2019. Cyberbullying detection on social multimedia using soft computing techniques: a meta-analysis. *Multimedia Tools and Applications* 78, 17 (2019), 23973–24010.
[85] Stephen S Leff. 2007. Bullying and peer victimization at school: Considerations and future directions. *School Psychology Review* 36, 3 (2007), 406–412.
[86] Suzet Tanya Lereya, Catherine Winsper, Jon Heron, Glyn Lewis, David Gunnell, Helen L Fisher, and Dieter Wolke. 2013. Being bullied during childhood and the prospective pathways to self-harm in late adolescence. *Journal of the American Academy of Child & Adolescent Psychiatry* 52, 6 (2013), 608–618.
[87] Melvin J Lerner and Dale T Miller. 1978. Just world research and the attribution process: Looking back and ahead. *Psychological bulletin* 85, 5 (1978), 1030.
[88] Qing Li. 2008. A cross-cultural comparison of adolescents' experience related to cyberbullying. *Educational Research* 50, 3 (2008), 223–234.
[89] Ziyi Li, Junpei Kawamoto, Yaokai Feng, and Kouichi Sakurai. 2016. Cyberbullying detection using parent-child relationship between comments. In *ACM International Conference Proceeding Series*. Association for Computing Machinery, 325–334. https://doi.org/10.1145/3011141.3011182
[90] Paul Benjamin Lowry, Jun Zhang, Chuang Wang, and Mikko Siponen. 2016. Why do adults engage in cyberbullying on social media? An integration of online disinhibition and deindividuation effects with the social structure and social learning model. *Information Systems Research* 27, 4 (2016), 962–986.
[91] Thabo Mahlangu, Chunling Tu, and Pius Owolawi. 2018. A review of automated detection methods for cyberbullying. In *2018 International Conference on Intelligent and Innovative Computing Applications (ICONIC)*. IEEE, 1–5.
[92] Amrita Mangaonkar, Allenoush Hayrapetian, and Rajeev Raje. 2015. Collaborative detection of cyberbullying behavior in Twitter data. In *IEEE International Conference on Electro Information Technology*, Vol. 2015-June. IEEE Computer Society, 611–616. https://doi.org/10.1109/EIT.2015.7293405
[93] Tolba Marwa, Ouadfel Salima, and Meshoul Souham. 2018. Deep learning for online harassment detection in tweets. In *Proceedings - PAIS 2018: International Conference on Pattern Analysis and Intelligent Systems*. Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/PAIS.2018.8598530
[94] Ersilia Menesini, Annalaura Nocentini, Benedetta Emanuela Palladino, Ann Frisén, Sofia Berne, Rosario Ortega-Ruiz, Juan Calmaestra, Herbert Scheithauer, Anja Schultze-Krumbholz, Piret Luik, et al. 2012. Cyberbullying definition among adolescents: A comparison across six European countries. *Cyberpsychology, Behavior, and Social Networking* 15, 9 (2012), 455–463.
[95] Diana J Meter, Ross Budziszewski, Abigail Phillips, and Troy E Beckert. 2021. A Qualitative Exploration of College StudentsâĂŹ Perceptions of Cyberbullying. *TechTrends* (2021), 1–9.
[96] Jerold D Miller and Shirley M Hufstedler. 2009. Cyberbullying Knows No Borders. *Australian Teacher Education Association* (2009).
[97] Tijana Milosevic. 2016. Social media companies' cyberbullying policies. *International Journal of Communication* 10 (2016), 22.
[98] 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.
[99] Farman A Moayed, Nancy Daraiseh, Richard Shell, and Sam Salem. 2006. Workplace bullying: a systematic review of risk factors and outcomes. *Theoretical Issues in Ergonomics Science* 7, 3 (2006), 311–327.
[100] R Morris, Daniel McDuff, and R Calvo. 2014. Crowdsourcing techniques for affective computing. In *The Oxford handbook of affective computing*. Oxford Univ. Press, 384–394.
[101] Michael Muller, Cecilia Aragon, Shion Guha, Marina Kogan, Gina Neff, Cathrine Seidelin, Katie Shilton, and Anissa Tanweer. 2020. Interrogating Data Science. In *Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing*. 467–473.
[102] Michael J Muller and Allison Druin. 2012. Participatory design: The third space in human-computer interaction. *The human-computer interaction handbook: Fundamentals, evolving technologies, and emerging applications* (2012), 1125–1154.
[103] Amgad Muneer and Suliman Mohamed Fati. 2020. A Comparative Analysis of Machine Learning Techniques for Cyberbullying Detection on Twitter. *Future Internet* 12, 11 (2020), 187.

[104] David Nabuzoka. 2003. Experiences of bullying-related behaviours by English and Zambian pupils: a comparative study. *Educational Research* 45, 1 (2003), 95–109.

[105] Samaneh Nadali, Masrah Azrifah Azmi Murad, Nurfadhlina Mohamad Sharef, Aida Mustapha, and Somayeh Shojaee. 2014. A review of cyberbullying detection: An overview. In *International Conference on Intelligent Systems Design and Applications, ISDA*. IEEE Computer Society, 325–330. https://doi.org/10.1109/ISDA.2013.6920758

[106] Vinita Nahar, Sanad Al-Maskari, Xue Li, and Chaoyi Pang. 2014. Semi-supervised learning for cyberbullying detection in social networks. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 8506 LNCS. Springer Verlag, 160–171. https://doi.org/10.1007/978-3-319-08608-8{_}14

[107] Vinita Nahar, Xue Li, Chaoyi Pang, and Yang Zhang. 2013. Cyberbullying Detection based on text-stream classification. In *Conferences in Research and Practice in Information Technology Series*, Vol. 146. 49–58. http://www.noswearing.com/

[108] Vinita Nahar, Sayan Unankard, Xue Li, and Chaoyi Pang. 2012. Sentiment analysis for effective detection of cyber bullying. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 7235 LNCS. Springer, Berlin, Heidelberg, 767–774. https://doi.org/10.1007/978-3-642-29253-8_75

[109] K. Nalini and L. Jaba Sheela. 2015. Classification of tweets using text classifier to detect cyber bullying. In *Advances in Intelligent Systems and Computing*, Vol. 338. Springer Verlag, 637–645. https://doi.org/10.1007/978-3-319-13731-5{_}69

[110] K. Nalini and L. Jaba Sheela. 2016. Classification using Latent Dirichlet Allocation with Naive Bayes Classifier to detect Cyber Bullying in Twitter. *Indian Journal of Science and Technology* 9, 28 (7 2016).

[111] Charisse L Nixon. 2014. Current perspectives: the impact of cyberbullying on adolescent health. *Adolescent health, medicine and therapeutics* 5 (2014), 143.

[112] William S Noble. 2006. What is a support vector machine? *Nature biotechnology* 24, 12 (2006), 1565–1567.

[113] Chitu Okoli and Kira Schabram. 2010. A guide to conducting a systematic literature review of information systems research. (2010).

[114] Scott W O'Leary-Kelly and Robert J Vokurka. 1998. The empirical assessment of construct validity. *Journal of operations management* 16, 4 (1998), 387–405.

[115] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kiciman. 2019. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data* 2 (2019), 13.

[116] Dan Olweus. 1994. Bullying at school. In *Aggressive behavior*. Springer, 97–130.

[117] Mahesh Pal. 2005. Random forest classifier for remote sensing classification. *International journal of remote sensing* 26, 1 (2005), 217–222.

[118] Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing Gender Bias in Abusive Language Detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 2799–2804. https://doi.org/10.18653/v1/D18-1302

[119] John Pavlopoulos, Jeffrey Sorensen, Lucas Dixon, Nithum Thain, and Ion Androutsopoulos. 2020. Toxicity Detection: Does Context Really Matter? *arXiv preprint arXiv:2006.00998* (2020).

[120] 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.

[121] Nektaria Potha and Manolis Maragoudakis. 2015. Cyberbullying detection using time series modeling. In *IEEE International Conference on Data Mining Workshops, ICDMW*, Vol. 2015-Janua. IEEE Computer Society, 373–382. https://doi.org/10.1109/ICDMW.2014.170

[122] Yada Pruksachatkun, Sachin R Pendse, and Amit Sharma. 2019. Moments of change: Analyzing peer-based cognitive support in online mental health forums. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.

[123] Rahat Ibn Rafiq, Homa Hosseinmardi, Richard Han, Qin Lv, and Shivakant Mishra. 2018. Scalable and timely detection of cyberbullying in online social networks. In *Proceedings of the ACM Symposium on Applied Computing*. Association for Computing Machinery, 1738–1747. https://doi.org/10.1145/3167132.3167317

[124] Elaheh Raisi and Bert Huang. 2016. Cyberbullying Identification Using Participant-Vocabulary Consistency. (6 2016). http://arxiv.org/abs/1606.08084

[125] Elaheh Raisi and Bert Huang. 2017. Cyberbullying detection with weakly supervised machine learning. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2017*. Association for Computing Machinery, Inc, New York, New York, USA, 409–416. https://doi.org/10.1145/3110025.3110049

[126] Elaheh Raisi and Bert Huang. 2019. Reduced-bias co-trained ensembles for weakly supervised cyberbullying detection. In *International Conference on Computational Data and Social Networks*. Springer, 293–306.

[127] Gonzalo Ramos, Jina Suh, Soroush Ghorashi, Christopher Meek, Richard Banks, Saleema Amershi, Rebecca Fiebrink, Alison Smith-Renner, and Gagan Bansal. 2019. Emerging Perspectives in Human-Centered Machine Learning. In

Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems. 1–8.

[128] Johan Redström. 2006. Towards user design? On the shift from object to user as the subject of design. *Design studies* 27, 2 (2006), 123–139.

[129] Kelly Reynolds, April Kontostathis, and Lynne Edwards. 2011. Using machine learning to detect cyberbullying. In *Proceedings - 10th International Conference on Machine Learning and Applications, ICMLA 2011*, Vol. 2. 241–244. https://doi.org/10.1109/ICMLA.2011.152

[130] Irina Rish et al. 2001. An empirical study of the naive Bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, Vol. 3. 41–46.

[131] Walisa Romsaiyud, Kodchakorn Na Nakornphanom, Pimpaka Prasertsilp, Piyaporn Nurarak, and Pirom Konglerd. 2017. Automated cyberbullying detection using clustering appearance patterns. In *2017 9th International Conference on Knowledge and Smart Technology: Crunching Information of Everything, KST 2017*. Institute of Electrical and Electronics Engineers Inc., 242–247. https://doi.org/10.1109/KST.2017.7886127

[132] Hugo Rosa, N Pereira, Ricardo Ribeiro, Paula Costa Ferreira, João Paulo Carvalho, Sofia Oliveira, Luísa Coheur, Paula Paulino, AM Veiga Simão, and Isabel Trancoso. 2019. Automatic cyberbullying detection: A systematic review. *Computers in Human Behavior* 93 (2019), 333–345.

[133] Derek Ruths and Jürgen Pfeffer. 2014. Social media for large studies of behavior. *Science* 346, 6213 (2014), 1063–1064.

[134] Koustuv Saha, Eshwar Chandrasekharan, and Munmun De Choudhury. 2019. Prevalence and psychological effects of hateful speech in online college communities. In *Proceedings of the 10th ACM Conference on Web Science*. 255–264.

[135] Semiu Salawu, Yulan He, and Joanna Lumsden. 2017. Approaches to automated detection of cyberbullying: A survey. *IEEE Transactions on Affective Computing* (2017).

[136] Nithya Sambasivan, Erin Arnesen, Ben Hutchinson, and Vinodkumar Prabhakaran. 2020. Fairness Considered Harmful: On the Non-portability of Fair-ML in India. *arXiv preprint arXiv:2012.03659* (2020).

[137] Huascar Sanchez and Shreyas Kumar. [n.d.]. *Twitter Bullying Detection.* Technical Report. https://www.researchgate.net/publication/267823748

[138] Martin Shepperd, David Bowes, and Tracy Hall. 2014. Researcher bias: The use of machine learning in software defect prediction. *IEEE Transactions on Software Engineering* 40, 6 (2014), 603–616.

[139] Shruthi and Prof Mangala C. 2017. A Framework for Automatic Detection and Prevention of Cyberbullying in Social Media. *International Journal of Innovative Research in Computer and Communication Engineering* 5, 6 (2017), 86–90. www.ijircce.com

[140] Jim Sidanius and Felicia Pratto. 2001. *Social dominance: An intergroup theory of social hierarchy and oppression*. Cambridge University Press.

[141] Theodore M Singelis. 1994. The measurement of independent and interdependent self-construals. *Personality and social psychology bulletin* 20, 5 (1994), 580–591.

[142] Vivek K. Singh, Souvick Ghosh, and Christin Jose. 2017. Toward multimodal cyberbullying detection. In *Conference on Human Factors in Computing Systems - Proceedings*, Vol. Part F1276. Association for Computing Machinery, 2090–2099. https://doi.org/10.1145/3027063.3053169

[143] Vivek K Singh and Connor Hofenbitzer. 2019. Fairness across network positions in cyberbullying detection algorithms. In *2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 557–559.

[144] Vivek K. Singh, Qianjia Huang, and Pradeep K. Atrey. 2016. Cyberbullying detection using probabilistic socio-textual information fusion. In *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2016*. Institute of Electrical and Electronics Engineers Inc., 884–887. https://doi.org/10.1109/ASONAM.2016.7752342

[145] Robert Slonje, Peter K Smith, and Ann Frisén. 2013. The nature of cyberbullying, and strategies for prevention. *Computers in human behavior* 29, 1 (2013), 26–32.

[146] Peter K. Smith, Jess Mahdavi, Manuel Carvalho, Sonja Fisher, Shanette Russell, and Neil Tippett. 2008. Cyberbullying: its nature and impact in secondary school pupils. *Journal of Child Psychology and Psychiatry* 49, 4 (4 2008), 376–385. https://doi.org/10.1111/j.1469-7610.2007.01846.x

[147] A. Squicciarini, S. Rajtmajer, Y. Liu, and C. Griffin. 2015. Identification and characterization of cyberbullying dynamics in an online social network. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2015*. Association for Computing Machinery, Inc, New York, New York, USA, 280–285. https://doi.org/10.1145/2808797.2809398

[148] B. Sri Nandhini and J. I. Sheeba. 2015. Cyberbullying detection and classification using information retrieval algorithm. In *ACM International Conference Proceeding Series*, Vol. 06-07-March-2015. Association for Computing Machinery. https://doi.org/10.1145/2743065.2743085

[149] B. Srinandhini and J. I. Sheeba. 2015. Online social network bullying detection using intelligence techniques. *Procedia Computer Science* 45, C (2015), 485–492. https://doi.org/10.1016/j.procs.2015.03.085

[150] Rekha Sugandhi, Anurag Pande, Siddhant Chawla, Abhishek Agrawal, and Husen Bhagat. 2015. Methods for detection of cyberbullying: A survey. In *2015 15th International Conference on Intelligent Systems Design and Applications (ISDA)*. IEEE, 173–177.

[151] Harini Suresh and John V Guttag. 2019. A framework for understanding unintended consequences of machine learning. *ArXiv* abs/1901.10002 (2019).

[152] Nargess Tahmasbi and Elham Rastegari. 2018. A socio-contextual approach in automated detection of public cyberbullying on Twitter. *ACM Transactions on Social Computing* 1, 4 (2018), 1–22.

[153] Hannah J Thomas, Jason P Connor, and James G Scott. 2015. Integrating traditional bullying and cyberbullying: challenges of definition and measurement in adolescents–a review. *Educational psychology review* 27, 1 (2015), 135–152.

[154] Robert S Tokunaga. 2010. Following you home from school: A critical review and synthesis of research on cyberbullying victimization. *Computers in human behavior* 26, 3 (2010), 277–287.

[155] Sabina Tomkins, Lise Getoor, Yunfei Chen, and Yi Zhang. 2018. A socio-linguistic model for cyberbullying detection. In *Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2018*. Institute of Electrical and Electronics Engineers Inc., 53–60. https://doi.org/10.1109/ASONAM.2018.8508294

[156] Laura P. v Bosque and Sara Elena Garza. 2014. Aggressive text detection for cyberbullying. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 8856 (11 2014), 221–232. https://doi.org/10.1007/978-3-319-13647-9_21

[157] Katrien Van Cleemput, Heidi Vandebosch, and Sara Pabian. 2014. Personal characteristics and contextual factors that determine "helping,""joining in," and "doing nothing" when witnessing cyberbullying. *Aggressive behavior* 40, 5 (2014), 383–396.

[158] Cynthia Van Hee, Gilles Jacobs, Chris Emmery, Bart DeSmet, Els Lefever, Ben Verhoeven, Guy De Pauw, Walter Daelemans, and VÃĪronique Hoste. 2018. Automatic detection of cyberbullying in social media text. *PLoS ONE* 13, 10 (10 2018). https://doi.org/10.1371/journal.pone.0203794

[159] Kathleen Van Royen, Karolien Poels, Walter Daelemans, and Heidi Vandebosch. 2015. Automatic monitoring of cyberbullying on social networking sites: From technological feasibility to desirability. *Telematics and Informatics* 32, 1 (2015), 89–97.

[160] Kiri Wagstaff. 2012. Machine Learning that Matters. *Proceedings of 29th International Conference on Machine Learning*. 1 (06 2012).

[161] Thilini Wijesiriwardene, Hale Inan, Ugur Kursuncu, Manas Gaur, Valerie L Shalin, Krishnaprasad Thirunarayan, Amit Sheth, and I Budak Arpinar. 2020. ALONE: A Dataset for Toxic Behavior Among Adolescents on Twitter. In *International Conference on Social Informatics*. Springer, 427–439.

[162] Dieter Wolke, Andrea Schreier, Mary C Zanarini, and Catherine Winsper. 2012. Bullied by peers in childhood and borderline personality symptoms at 11 years of age: a prospective study. *Journal of child psychology and psychiatry* 53, 8 (2012), 846–855.

[163] Richmond Y Wong and Vera Khovanskaya. 2018. Speculative Design in HCI: From Corporate Imaginations to Critical Orientations. In *New Directions in Third Wave Human-Computer Interaction: Volume 2-Methodologies*. Springer, 175–202.

[164] Mengfan Yao. 2019. Robust detection of cyberbullying in social media. In *The Web Conference 2019 - Companion of the World Wide Web Conference, WWW 2019*. Association for Computing Machinery, Inc, 61–66. https://doi.org/10.1145/3308560.3314196

[165] Mengfan Yao, Charalampos Chelmis, and Daphney Stavroula Zois. 2018. Cyberbullying detection on instagram with optimal online feature selection. In *Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2018*. Institute of Electrical and Electronics Engineers Inc., 401–408. https://doi.org/10.1109/ASONAM.2018.8508329

[166] Mengfan Yao, Charalampos Chelmis, and Daphney Stavroula Zois. 2019. Cyberbullying ends here: Towards robust detection of cyberbullying in social media. In *The World Wide Web Conference*. 3427–3433.

[167] Xiang Zhang, Jonathan Tong, Nishant Vishwamitra, Elizabeth Whittaker, Joseph P. Mazer, Robin Kowalski, Hongxin Hu, Feng Luo, Jamie Macbeth, and Edward Dillon. 2016. Cyberbullying Detection with a Pronunciation Based Convolutional Neural Network. In *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*. Institute of Electrical and Electronics Engineers (IEEE), 740–745. https://doi.org/10.1109/icmla.2016.0132

[168] Rui Zhao and Kezhi Mao. 2016. Cyberbullying detection based on semantic-enhanced marginalized denoising auto-encoder. *IEEE Transactions on Affective Computing* 8, 3 (2016), 328–339.

[169] Rui Zhao, Anna Zhou, and Kezhi Mao. 2016. Automatic detection of cyberbullying on social networks based on bullying features. In *ACM International Conference Proceeding Series*, Vol. 04-07-Janu. Association for Computing Machinery, New York, New York, USA, 1–6. https://doi.org/10.1145/2833312.2849567

[170] Haoti Zhong, Hao Li, Anna Cinzia Squicciarini, Sarah Michele Rajtmajer, Christopher Griffin, David J Miller, and Cornelia Caragea. 2016. Content-Driven Detection of Cyberbullying on the Instagram Social Network.. In *IJCAI*. 3952–3958.
[171] Haoti Zhong, David J Miller, and Anna Squicciarini. 2018. Flexible Inference for Cyberbully Incident Detection. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 356–371.
[172] Karolina Zwierzynska, Dieter Wolke, and Tanya S Lereya. 2013. Peer victimization in childhood and internalizing problems in adolescence: a prospective longitudinal study. *Journal of abnormal child psychology* 41, 2 (2013), 309–323.

Received January 2021 ; revised April 2021 ; accepted May 2021