IJORP
International Journal of Research and Publications

CYBERBULLYING DETECTION USING HUMAN-CENTEREDNESS ALGORITHM

Chukwudi-Osondu Tochukwu O.
Tochukwu2550@gmail.com
Computer Science Department, Federal Polytechnic Oko, Anambra State.

Abstracts

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 on the roles, beliefs, desires, or expectations of humans.

This study analyzed few papers based on a three-prong human-centeredness algorithm design framework – spanning theoretical, participatory, and speculative design. It was 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, 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 were discussed.

Key Words*: Cyberbullying detection, human-centered machine learning, Human-centered computing, literature review, social media*

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 and has been a pervasive problem in the society for several decades. With the proliferation of digital social technologies among teens and young adults, bullying, once restricted to the school or neighborhood, has now moved into the digital realm.

Cyberbullying is a form of bullying or harassment using electronic means which includes sending, posting or sharing negative, harmful, false or mean content about someone else. It can take place on social media, messaging platforms, gaming platforms and mobile phones. Cyberbullying inflicts unforgettable pain on the victims, with close to two-thirds of adolescents already having experienced some form of cyberbullying ranging from offensive name-calling to spreading of false rumors. Mental health issues such as anxiety and depression are known to be a result of experiencing bullying as children. The trauma from bullying can also lead to increased suicidal ideation and self-harm . Being bullied at the start of the teenage years has been shown as a potential indicator of the disposition towards borderline personality disorder symptoms . Given its prevalence and long-lasting damage inflicted on young victims of bullying, 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 .

However, the massive volumes of evolving, real-time, multimodal, heterogeneous, and unstructured social media data makes manual detection of cyberbullying intractable. 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. In addition the detection mechanisms can also provide support to the victims along with ways to identifying the perpetrators.

LITERATURE REVIEW

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.

However, cyberbullying detection is not merely a classification task to identify which and whose content might be abusive towards an individual or group. Reasons range from the sensitivities around the cyberbullying experience, its effects on the victim(s) and bystander(s) , social stigma, impact on the health and functioning of online communities, to the potential far-reaching ramifications of cyberbullying incidents on various stakeholders.

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, what signals in online data serve as the most salient features in classification, what types of machine learning methodologies have been adopted, how the performances of different models and datasets compare against one another based on standardized metrics like accuracy, precision, and recall, and how the paucity of standardized datasets and reliance on manual annotation has hampered reproducibility and replicability.

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 (2019) 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., 2018) on this topic. Other scholars noted that high quality datasets are lacking in this area, 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 (Al-Garadi et al, 2019). Vast majority have relied on public social media data, 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. Emmery et al. (2020) 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.

Importantly, due to the inherently subjective interpretation and experience of cyberbullying incidents, researchers have argued that human annotators, used for training data generation, may have different views on which sample is passed as cyberbullying. Subjectivity is not just limited to training data duration; it may exist during the creation of a set of features as well – a fact argued by 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 (Dadvar et al., 2013). 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. Consequently, (Rosa et al., 2019), after a systematic review of 22 papers, advocated for establishing well-defined criteria that could help duration 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.

METHODOLOGY

A second, complementary set of reviews have focused on the underlying machine learning methodology in cyberbullying detection. A notable survey of prior research by (Salawu et al., 2017) 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 (2014) 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., 2016) emphasized, "*most of these [cyberbullying] studies haveglossed over the central issue: the role of ... social media artifacts themselves in promoting cyberbullying*." Al-Garadi et al (2019) 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, 2010) 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, including unsupervised and semi-supervised techniques (Emmery et al.,(2020). 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 . Therefore, researchers have valued careful selection of an evaluation metric that is independent of data skewness, to avoid uncertain results and undesirable outcomes (Emmery et al., 2020). 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., 2019), 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.

A Human-Centered Perspective of Machine Learning

Machine learning is increasingly adopted to address societal problems via data-driven decisionmaking (Chancellor et al., 2019), 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.

First, humancenteredness, 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 (Baumer, 2017). For cyberbullying detection research, these theories can be incredibly valuable – many rich psychological theories like the Control balance theory, Dominance theory, Just world belief , and Crime opportunity theory 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.

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 (Baumer, 2017).

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. 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., 2019), 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.

A Human-Centered Algorithm Design Framework

Baumer (2017) conceptualized human-centered algorithm design to engender three key dimensionsor strategies – *theoretical, participatory, and speculative design*. These dimensions are neither sequential nor mutually exclusive, but rather, "provide a sense for the range of possibilities". 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:

- ***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 . 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 (Chancellor et al., 2019), , aiding them in their dataset selection, feature selection, and model evaluations.
- ***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. 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. 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.

- ***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 (Auger, 2013). 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.

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 into the design of technology, and accounting for stakeholders in the creation of technological solutions" (Chancellor et al., 2019), . Scholars in the evolving and emergent area of humancentered machine learning have therefore argued that machine learning needs to stay grounded in human needs (Chancellor et al., 2019), , models need to be built in inclusive ways that adequately represent the diverse experiences of different individuals and minimize biases, and that machine learning approaches ought to incorporate interpretability and transparency to not only elucidate its potential for harm, but also how data-driven decisions are used in practical scenarios. 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. Amershi et al (2014) 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. Articulating these roles and representing them in the development of machine learning algorithms can point to differing agencies between people and the algorithms.

Adopting the threeprong human-centered algorithm design lens proposed by (Baumer, 2017), 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.

DISCUSSION AND FINDINGS

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. But the algorithmic transformation of theoretical concepts, as is the case for cyberbullying complicates opportunities for theoretical hypothesis testing (Baumer, 2017), 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.

- ***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 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 cyberbullying detection techniques developed in an a theoretical 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 whichplatform, and structural idiosyncrasies stemming from

different platforms' distinct characteristics.

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.

- ***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, and therefore, like other real-world problems, 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. 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?

- ***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. 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. 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.

CONCLUSION

After establishing a corpus of relevant documents to cyberbullying detection, this analyzed the human involvement in the development of these models using an established human-centered algorithmic design framework (Baumer, 2017). This paper specifically reviewed the past research in terms of their considerations for theoretical, participatory, and speculative design. The 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 duration 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.

REFERENCES

Akshi, k., &Nitin, S. (2019) Cyberbullying detection on social multimedia using soft computing techniques: a meta-analysis.

Thabo, M., Chunling T., & Pius O. (2018) A review of automated detection methods for cyberbullying.

Mohammed A., Mohammad R., Nawsher K., Ghulam M., Henry N., Ihsan A., Ghulam M., Haruna Ch., Hasan K., & Abdullah G. (2019) Predicting cyberbullying on social media in the big data era using machine learning algorithms: review of literature and open challenges.

Aboud S., Nadine F., & Gerard M. (2020) A review on non-supervised approaches for cyberbullying detection. *International Journal of Engineering Pedagogy* 10.

Chris E., Ben V., Guy P., Gilles J., Cynthia H., Els L., Bart D., Véronique H., & Walter D. (2020) Current limitations in cyberbullying detection: On evaluation criteria, reproducibility, and data

scarcity. *Language Resources and Evaluation*.

Maral D., Dolf T., Roeland O., &Franciska J. (2013) Improving cyberbullying detectionwith user context. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence andLecture Notes in Bioinformatics*, Berlin, Heidelberg, https://doi.org/10.1007/9783-642-36973-5_62

Hugo R., Pereira N., Ricardo R., Paula F., &João C (2019) Automatic cyberbullying detection: A systematic review.

Saleema A., Maya C., William B., &Todd K. (2014) Power to the people: The role of humans in interactive machine learning. *Ai Magazine*35, 4, 105–120.

James A. (2013) Speculative design: crafting the speculation. *Digital Creativity*24, 1 (2013), 11–35.

Baumer, E. (2017)Toward human-centered algorithm design. *Big Data & Society* 4, 2.

Stevie, C., Baumer, E., &Munmun C. (2019) Who is the "Human" in Human-Centered Machine Learning: The Case of Predicting Mental Health from Social Media.