"A Tale on Abuse and Its Detection over Online Platforms, Especially over Emails": From the Context of Bangladesh

Ishita Haque
1017052031@grad.cse.buet.ac.bd
Bangladesh University of Engineering and Technology
Dhaka, Bangladesh

Rudaiba Adnin
1505032.ra@ugrad.cse.buet.ac.bd
Bangladesh University of Engineering and Technology
Dhaka, Bangladesh

Sadia Afroz
1505030.sa@ugrad.cse.buet.ac.bd
Bangladesh University of Engineering and Technology
Dhaka, Bangladesh

Faria Huq
1505052.fh@ugrad.cse.buet.ac.bd
Bangladesh University of Engineering and Technology
Dhaka, Bangladesh

Sazan Mahbub
1505020.sm@ugrad.cse.buet.ac.bd
Bangladesh University of Engineering and Technology
Dhaka, Bangladesh

A. B. M. Alim Al Islam
razi_bd@yahoo.com
Bangladesh University of Engineering and Technology
Dhaka, Bangladesh

ABSTRACT

With the advent of pervasive usage of online platforms, online abusive behavior has become an indispensable part of our life demanding great attention from the research community. Accordingly, the research community is spending its effort on the demanding task, however, perhaps having much less effort on emails, even though emails are identified as a prominent source of exchanging online abusive behaviors. To fill in this gap in the literature, we conduct an in-depth study to investigate online abusive behavior having a special focus on emails. To do so, we perform a mixed-method user study consisting of formative interviews (n=15) and a survey (n=65) over user's experience, coping strategies, etc., pertinent to online abuse in Bangladesh, especially focusing on abuse over emails. We also dig into users' perspectives to analyze strengths and challenges associated with different types of abuse detection systems for online platforms, especially for emails. One of the noteworthy findings of our study is that there exists a significant demand for abuse detection systems over emailing platforms even after having a lesser frequency of abuse occurring over emails. Our findings also highlight a certain level of user preference for an automated abuse detection system potentially considering its more control and fewer privacy concerns to users, however, being challenged due to having the limitation of lesser ability to detect implicit abuse. We also identify several limiting factors associated with a human-moderator-based abuse detection system, including less comfort, less trust in different types of moderators, inhumane demands to the moderators, and time delay in detecting abuses. These findings point to opportunities for design interventions for hybrid abuse detection systems, which is the most preferred system to the users, to overcome all the limitations of automated and human-moderator-based systems.

CCS CONCEPTS

• **Online abuse → Abuse detection systems**; *Privacy*; Preventive mechanisms; • **Human centered computing**;

KEYWORDS

Online abuse, Abuse detection systems, Privacy, Human-centered computing, Emails, User studies

ACM Reference Format:
Ishita Haque, Rudaiba Adnin, Sadia Afroz, Faria Huq, Sazan Mahbub, and A. B. M. Alim Al Islam. 2021. "A Tale on Abuse and Its Detection over Online Platforms, Especially over Emails": From the Context of Bangladesh. In *8th International Conference on Networking, Systems and Security (8th NSysS 2021), December 21–23, 2021, Cox's Bazar, Bangladesh.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3491371.3491374

1 INTRODUCTION

With the rise of interactive online platforms, online abuse is becoming more and more prevalent. According to a recent report, about 47% of internet users reported being victims of some form of online harassment or cyberbullying [26]. Such experiences can be detrimental to the physical and mental well-being of the victims and cause anxiety, depression, low confidence, self-harm, low self-esteem, rejection by peers, aggression, hopelessness and even suicide contemplation [41], [46]. Victims tend to respond to abusive behaviors in a limited number of ways such as blocking and reporting abusive content. Victims even go to extreme lengths such as changing email addresses and deleting social media accounts to save themselves from online abuse [19], [41]. Some victims tend to avoid tech in general from the trauma created by online abuse [19].

Detection and moderation of the online have become very crucial, and therefore, have attracted attention in both research and public discourse. Social platforms such as Facebook, Twitter, Reddit, Pinterest, Tumblr, Instagram, etc., have their policies to define abusive behaviors and to decide on platforms' appropriate responses to them. Platform responses include restricting or disabling abuser accounts, sending warnings to potential abusers, removing abusive

8th NSysS 2021, December 21–23, 2021, Cox's Bazar, Bangladesh

ACM ISBN 978-1-4503-8737-8/21/12…$15.00
https://doi.org/10.1145/3491371.3491374

contents, etc. [34], [17], [44]. Online platforms have incorporated different approaches to moderate abusive content based on their operational policies. Existing approaches can be automated, human moderated, or can incorporate both automation and human moderation [9], [8]. Human moderators regulate the abusive contents and verify the contents reported by the users before taking any preventive measure [48], [17]. On the other hand, automated systems may incorporate machine learning-based approaches for detecting abusive content. Several platforms such as Facebook [17], Twitter [38], and Youtube [48] incorporate both human moderation and automated moderation to detect abusive behaviors. Some of these platforms provide system-generated assistance to the human moderators [8]. Even with such assistance , the human moderators are burdened with the extreme workload and detrimental effects on mental health because of their constant exposures to abusive contents [14], [31], [37]. Moreover, several platforms have been criticized for their inaccurate detection of online abusive cases [22], [32], as well as a delayed response against such cases.

Despite the growing interest in content moderation in the fields of Human-Computer Interaction (HCI) and Computer-Supported Cooperative Work (CSCW), most research on abuse detection and moderation have focused on abuse in social platforms [8], [5]. However, studies show that people also get substantially victimized by abusive behaviors over emails [35]. Known peers and online-only contacts are the sources of such common abusive behaviors [46]. These sources can target their victims over emailing platforms and social network sites, using them as a media of communication [35]. Moreover, abusive behavior detection systems and preventive mechanisms might not be similarly applicable for both - emails and social networks - owing to their inherent dissimilarities between their own operational modalities. Even after having this reality, abusive behavior detection systems and preventive mechanisms for emailing platforms are rarely explored. The existing approaches for detecting abusive behavior in emailing platforms incorporate human moderators, and this involvement can impose privacy concerns to the users as well as cause frustration with time delay [28] in delivering important emails. Even appointing acquaintances such as friends as moderators may not be convenient. It can cause discomfort to the victim, overburden moderators, and create the possibility of secondary trauma [6].

We explore the following research questions from the context of Bangladesh.

- **RQ1:** What are the effects of experiencing abusive behaviors over online platforms? How do users cope with such abusive behaviors?
- **RQ2:** How do users perceive abuse detection systems over online platforms, especially over emailing platforms?
- **RQ3:** What are the strengths and challenges users identify while using existing types of abuse detection systems, especially for emailing platforms?
- **RQ4:** What are the desirable design components in an abuse detection system, especially for emailing platforms to fulfill user needs?

Considering the above-mentioned aspects, in this study, we conduct a mixed-methods study that includes a semi-structured interview study with 15 participants and an anonymous online survey with 65 participants. The objective of this study to discover online abuse faced by people and how they use traditional abuse detection systems to address them.

1.1 Our Contributions

Through our extensive literature review and rigorous user studies, we make the following set of contributions in this paper.

- We present a detailed mixed-method study that examines the online abuse experiences of users from Bangladesh, the consequences they face due to abuse, and their coping practices which will allow the HCI community to better understand the problems associated with abuse detection.
- Through our user studies, we identify the perception that users of Bangladesh possess regarding current abuse detection systems, and present findings specific to the rarely explored domain of abuse over emails in Bangladesh.
- Through our user studies, we have identified the preference of users of Bangladesh among different types of abuse detection systems particularly for emailing platforms, while highlighting the strengths and limitations of each type from a user's perspective.
- We discuss the implications of our research findings to suggest possible design directions for overcoming the limitations over current abuse detection systems over emailing platforms. We further suggest areas worth exploring to HCI researchers in the domain of online abuse detection in the context of Bangladesh.

The remainder of the paper is structured as follows. Section 2 summarizes the key elements of related work. Section 3 provides relevant information on our methodology. Section 4 presents our research findings. Section 5 presents our discussion on our research findings. Section 6 shows the limitations we faced while doing the research. Section 7 gives pointers to the scope of further research, and finally, Section 8 concludes with a summary of contributions.

2 RELATED WORK

In this section, we discuss existing research on email-based and other online abuses and the limitations posed by current abuse detection systems.

Online abusive behaviors can come in many forms such as sexual harassment [13], cyber-stalking [2], cyber-bullying [11], flaming [41], threat [34], etc. A study conducted by Pew Research in 2017 reports that about 41% of Americans have been personally subjected to online harassment and about 66% have witnessed these behaviors directed at others [13]. Young adults, women [26], and those who identify as LGBTQ+ [19] are more likely to experience online abuse. The impacts of online abuse are multidimensional. Online abuse can cause anxiety, depression, low empathy, low self-esteem, declining confidence levels, and even suicide [33]. A Pew Research study found that 13% of adults in the United States had experienced mental or emotional stress as a result of online harassment [13]. According to a recent Data & Society Research Institute report, 27% of internet users self-censor their online postings out of fear of online harassment and 43% of victims of online abuse had to change their contact information to escape their abuse [26]. Research conducted on South-Asian women shows that only 1% of women sought out help from law-enforcement agencies [41].

Existing detection approaches can be broadly categorized as human-moderated and automated. Human-moderator-based detection systems can be further divided into two types, centralized and distributed [8]. The centralized approach includes content detection by teams of paid or unpaid moderators, volunteer or externally contracted organizations by the platform [18], [8]. In the distributed approach users down-vote the undesirable content, often reporting the content [24]. Sites such as Reddit, Stack Overflow and Yik Yak use distributed social detection and moderation [42]. Automated approaches include using machine learning-based models to detect abusive content [8]. A combination model is used often, where machine learning-based models can help in triaging content before human moderators review them [15].

Emails play a very important role in personal and professional communication [23]. In 2019, the number of global e-mail users amounted to 3.9 billion and is set to grow to 4.48 billion users in 2024 [10]. Emails have become a significant medium of online abuse as well. A study conducted by Pew Research shows that when participants were asked to recall where their most recent experience of online harassment took place, 16% said in a personal email account [12]. Although much research has been done on spam detection and prevention, very few have explored abuse over emails. A tool named SquadBox [28] helps users prevent abusive behavior by keeping a list of trusted friends, volunteers, or paid moderators between the world and the users' inbox. However, abusive content detection and moderation can potentially become an overload for any moderator, friends, or not, and exposure to such content can be detrimental to the well-being of the moderator [4]. Features like blocking and reporting allow users to defend themselves to some extent [16], [43], but do not allow users to prevent abusive behavior beforehand and the victim has already been impacted by then.

Social platforms such as Facebook, Youtube, and Twitter have incorporated commercial content moderation [45] and have incorporated both human moderation and automated detection to detect abusive content that violates its policies [15], [38]. Reddit has its policies about defining abusive behaviors and includes moderators who regulate content generated within subreddits voluntarily [20]. CrossMod, an automated abusive behavior detection system for Reddit, is built to assist human moderators, to lessen their workload [8]. BoC is a cross-platform automated abuse detection system that is built to allow communities to deal with abusive behaviors [9]. Tune is an experimental Chrome extension that detects abusive comments and lets people customize the level of toxicity they want to see in comments across the internet [1].

Numerous platforms depend on human moderators [37], however, involving human moderators poses some concerns. Perception of abuse to a human moderator may differ significantly from the user. Research on user experience with content moderation in Reddit reveals that about 18% of the participants expressed that their post removal was appropriate and about 29% of the participants expressed some level of frustration about the removal [21]. Often, gap in perception of abuse between moderator and victim exists because of social and language barriers [41]. Moreover, online content is ever-increasing, therefore, human-moderator-based systems are burdened to appoint more and more moderators continuously. Several reports show that moderators are burdened with an inhumane level of workload [36]. Each content receives attention for merely a few seconds [18]. As a result, users' expectations to have a fair judgment based on cultural sensitivity may not be guaranteed in such fast decisions. Users express privacy concerns with human moderators and prefer that sensitive information such as financial information should not be viewed by moderators, even if they are friends of them [28]. People are often hesitant to reveal abuse, except within their immediate support systems [41], therefore, the victim may feel exposed and vulnerable in a human-moderator-based abuse detection system. Human moderators face numerous mental health issues such as self-harm, depression, anxiety, etc. [3], [30]. The constant exposure to abusive content substantially affects the mental health of human moderators [40], [31], [47]. Moreover, for time-sensitive platforms such as emailing platforms, human moderation creates time delays, causing frustration among users.

3 METHODOLOGY

We framed our research from the perspective of users and their need for technological support while they face online abuse, especially over emailing platforms. We also attempted to explore the challenges of the current abuse detection systems mostly over emailing platforms. Subsequently, we want to propose the future potential scope of the abuse detection systems particularly over emailing platforms. In doing so, we conducted semi-structured interviews and an online survey. We performed interviews with 15 participants and an online survey of 65 participants where we discovered their experiences while facing online abuse. The timeline of our interviews with the participants was between May to June 2020. The survey was conducted between July to August 2020. The authors of this paper were born and raised in Bangladesh. Therefore, the social media and academic networks of the authors facilitated access to the interviewees and survey participants.

3.1 Research Ethics and Anonymization

The study and data collection were approved by the Ethics Committee, a part of Integrity Strategy and Innovation, of the institution of the authors. To guarantee ethical conduction of our research, we ensured the confidentiality and anonymity of the participants in our study. Throughout our user studies, we pursued making them feel comfortable by creating rapport and ensuring non-judgemental communication. Before recruitment, the participants were notified about the purpose of the study, the type of questions they would be asked, the data collection process, and the affiliations of the researchers. Before interviews, we sought verbal consent from each participant for audio recording and using the provided information for research purposes. Interview participants were also given the freedom to choose between a phone call and a one-to-one Zoom meetings. We stored collected data in a private Google Drive accessible to the authors only. Throughout this paper, we use pseudonyms to identify our participants to protect their privacy.

3.2 Interviews

We conducted a formative interview study to investigate people's experiences with online abuse and their thoughts regarding abuse detection systems, with a special interest on email-based abuses.

All participants are of Bangladeshi origin. Prior research shows that victims of online abuse in South-East Asia are uncomfortable in sharing their experiences with abuse, and mostly seek help from family and friends [41]. Therefore, we recruited our participants through personal networks. Participants who use emails on regular basis were chosen. The target participants are people of all ages. We maintained diversity in age and gender during recruitment. 15 interviewees participated in 20 minutes to 40 minutes long semi-structured interviews with three authors over phone calls and Zoom meetings over multiple sessions. We conduct the interviews in the local language, i.e., Bengali. The participants aged 18 to 46 years, with an average of 28.53. Out of 15 participants, 8 participants report being female and 7 participants report being male. Table 1 lists the demographics of the participants. Participants are presented throughout the paper with codes (P#) to protect their identities.

Demographic information was collected at the start of the interview. The interviews were initiated with questions about their familiarity and frequency of using emailing platforms. After developing some rapport with the participants, we asked them questions to understand their experience with abusive behavior over emails and other online platforms, the impacts such behaviors create on their lives, and their responses to them. We then followed by asking questions about their perception of abuse detection systems and their concerns associated with using them. The interview comprised 32 numeric, categorized, and open-ended questions. We translated and transcribed the audio recordings of the interviews. Four of the authors carefully studied the transcripts and independently performed open-coding and thematic analysis [7] on them. After sharing the codes between themselves, they developed high-level categories that outline the key themes, and clustered the data accordingly. Table 2 shows the themes and related descriptions.

3.3 Online Survey

For the survey, the target participants are people of all ages from Bangladesh. The survey was done by following non-probabilistic sampling and snowball sampling. We surveyed 65 people aged between 16 to 52 years. Among our 65 participants, 39 reported as males, and 26 as females.

The questionnaire of the survey was distributed using Google form which is primarily common among the participants. Survey questions include both open-ended and closed-ended questions related to experiencing and taking actions against abusive behavior, abuse detection systems over emails, concerns, and challenges associated with using it. Since we surveyed after taking formative interviews, the options for close-ended questions in the survey were placed with the help of responses of the interviews. There are 34 questions in the questionnaire. Some of the questions are *"How did the abusive comment(s) over email(s) affect you?", "Have you faced any abusive comment(s) on any other platform(s) except over email(s)?", "How did the abusive comment(s) on the online platform(s) affect you?", "What did you do when you faced any online abusive comment(s) in your daily life?", "Did you find any shortcomings of the existing abuse detection system(s)? Elaborate if you found any?", "What is the concern(s) you have with privacy issues in case of using a human moderator based abuse detection system?", "What is the concern(s) you have with privacy issues in case of using an automated abuse detection system in emailing platforms?", "What is the concern(s) you have with privacy issues in case of using an automated abuse detection system in emailing platforms?"*, etc. We analyze and compare survey results with the results from the interview study to gain further insights into needs and concerns related to abuse detection systems over emailing platforms.

4 RESEARCH FINDINGS

In the semi-structured interview study, we explore our research questions by asking probing questions to the participants to get rich insights on their experiences with abusive behaviors over emailing and other platforms and the actions they had taken to cope with such experiences. We deeply analyze the challenges the participants felt in existing types of abuse detection systems and obtained results on their preferences between these types and the factors influencing them. In the online survey, we explore our research questions further on a wider scale to gain results concerning the themes we developed. In this section, we discuss our findings associated with the themes (Table 2) developed from both user studies.

4.1 Effects of Online Abuse

Almost all of the participants in our study had experience facing abusive behavior on online platforms. Out of 15 participants of interviews, 12 participants (6 male and 6 female) reported facing abusive behavior online, which portrays the severity of the problem. 59% of the survey participants reported the same as well, which gives further credence to the extent of online abuse. Participants mentioned being abused in platforms such as Facebook, WhatsApp, Messenger, LinkedIn, Instagram, and over emails. P3 mentioned facing such abusive behaviors mostly over Facebook. He also mentioned one incident of being a victim by his friend, *"I have faced abusive comments over public posts on Facebook. In public groups, I have faced threatening and harassing comments. Once I had some issues with one of my friends and he opened a fake account with my name and wrote abusive comments. So everyone got very confused which profile was the real one."*

All the participants are familiar with emailing platforms and expressed emails as an important medium of communication in their lives. About 73% participants of the interview participants and about 57% of the survey participants mentioned using emailing platforms always. Despite using email frequently, participants mentioned fewer occurrences of abuse over emails. About 40% of the interview participants and 11% of the survey participants mentioned facing abusive behavior over emails.

The effects of online abuse ranged from mild to severe. Some of the participants of the interviews expressed they feel stressed, afraid, anxious, depressed, frustrated, angry, traumatized while facing abusive behavior in emailing and other online platforms. P7 mentioned, *"The emails were a bit threatening. Those affected me a lot. I needed much time to recover from that incident"*.

Some participants mentioned how the abusive behaviors greatly impacted their mental and social life, causing long-lasting trauma with any form of online and social interaction, and even forcing them into withdrawal. P2 mentioned, *"It affected me pretty badly. For 3-4 years I didn't use Facebook, email, or had any social media presence. I was very scared of any type of screens."*

Among the 38 survey participants who admitted facing online abusive behavior, 68.4% participants get stressed, 47.4% get anxious, 44.7% get depressed, 28.9% feel their confidence get low, and 2.6% do self-harm as a response to the abuse they experience online. Other

Table 1: Demographics of participants interviewed and their preference for the type of abuse detection system, especially over emailing platforms.

Participant Name	Occupation	Gender	Age	Preference in Type of Abuse Detection System, especially over Emailing Platforms
P1	IT Consultant	Female	24	Combination of automated and human moderated
P2	Student	Female	26	Combination of automated and human moderated
P3	Student	Male	18	Human-moderated
P4	Teacher	Female	44	Automated
P5	Banker	Male	33	Automated
P6	Software Engineer	Male	28	Combination of automated and human moderated
P7	Student	Female	26	Combination of automated and human moderated
P8	Student	Male	21	Combination of automated and human moderated
P9	Banker	Male	46	Automated
P10	Doctor	Female	36	Automated
P11	Student	Male	24	Combination of automated and human moderated
P12	Doctor	Female	24	Automated
P13	Student	Female	25	Automated
P14	Student	Female	22	Combination of automated and human moderated
P15	Job Holder	Male	31	Combination of automated and human moderated

Table 2: List of themes and descriptions associated with the interviews and the survey.

Theme	Description
Effects of online abuse	Experience and effects of online abuse on the users
Coping strategies against online abuse	Actions taken by users after experiencing online abuse
Perception about existing abuse detection systems, especially over emailing platforms	The perception of users about the existing abuse detection systems as well as their experiences with existing systems.
Strengths of existing types of abuse detection systems, especially for emailing platforms	The positive aspects identified by users while using different types of abuse detection systems, especially for emailing platforms
Challenges of existing types of abuse detection systems, especially for emailing platforms	Challenges faced while using existing abuse detection systems such as concerns related to privacy, preventive mechanisms, etc.
Expectations from an abuse detection system, especially for emailing platforms	Users' expectations and desired features of an abuse detection system

effects the participants feel include becoming aggressive, becoming hopeless, losing self-esteem, and their level of empathy going down. Figure 1 (a) demonstrates all the effects of online abuse mentioned by the victims among the survey participants.

4.2 Coping Strategies against Online Abuse

We discover the coping strategies of the participants when they face abusive behaviors, such as, blocking the abusive source, seeking friends' and relatives' help, often ignoring and even changing account credentials.

We observed that the majority of the participants preferred to solve situations by themselves before seeking any help. This observation is supported by the fact that blocking is the most popular response reported by the participants. 94% of the survey participants use the help of blocking or restricting the source of the abuse as a response to abuse. As mentioned by P2, *"I mainly block abusive accounts. If I see someone else getting harassed, I block that abuser as well."* However, a significant number of the participants expressed interest and trust in contacting trusted people, law enforcement agencies, or others in case of experiencing online abuse. 87% of the interview participants and 71% of the survey participants expressed a positive opinion in seeking help from them. Figure 2 demonstrates the comparison of the participants' trust to seek help.

Participants discuss their support system in facing online abusive behavior. Most participants mentioned seeking help from friends, family, and relatives before considering legal options. 17% of the survey participants mentioned contacting family members, friends, and relatives when they faced online abuse.

While some participants applauded the recent active initiatives of law enforcement agencies in Bangladesh and expressed interest

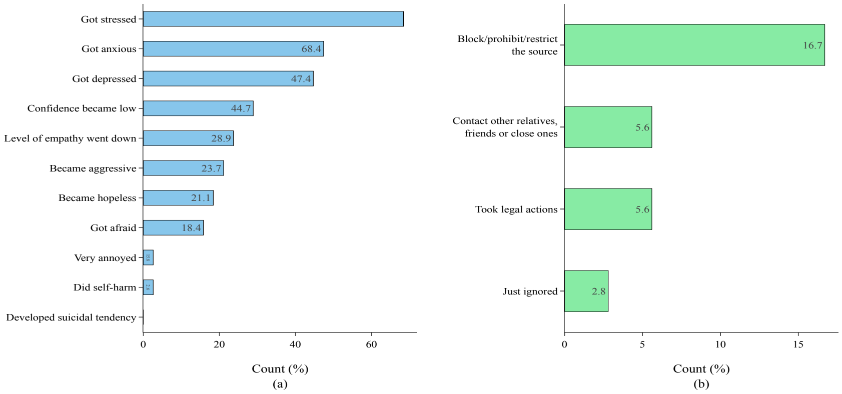

Figure 1: (a) Effects of online abuse (respondents=38), and (b) coping strategies (respondents=36) by the survey participants

to contact them, others mentioned less trust and comfort to contact law enforcement authorities to understand and be considerate of their personal stories of abuse. We observe that very few participants actually sought help from law enforcement agencies. Many participants expressed law enforcement authorities as the last resort, P2 mentions,*"At first I will try to resolve abusive incidents by consulting with my trusted people. If there is no progress, I will contact law enforcement agencies or police."* Only 1 among 15 interviewees actually took the step, but even he did not proceed further. Among the survey participants, only 5.6% mentioned taking legal actions.

Some participants mentioned ignoring the abuse as a coping mechanism as well. Among the survey participants, 2.8% mentioned opting to stay silent and ignore the incident of online abuse.

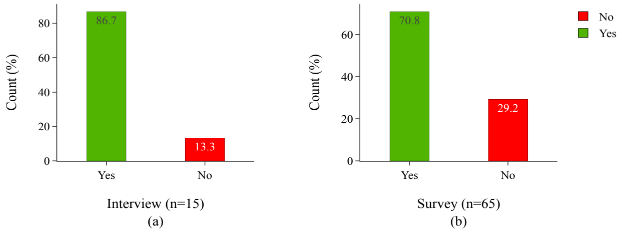

Figure 2: Comparison of participant's trust on others in case of facing online abuse in emails based on the question - *"Would you trust others to help you in case of facing online abuse in emails?"*

4.3 Perception about Existing Abuse Detection Systems, especially over Emailing Platforms

Regardless of the low frequency of abuse occurring through emails as reported by our participants, participants showed great interest in having an efficient abuse-detection system for emails. 13 of our 15 participants in interviews and 74% of our survey participants expressed such needs. While explaining this contrast between personal experience of abuse over emails and expectation for an abuse detection system for emails, participants mentioned being aware of such incidents occurring with someone else, which lead them to believe that such a system is needed for all. According to P1, *"It depends on the frequency of abusive emails that a person receives. Over the emailing platforms I used, I faced 2-3 such mails, so the frequency is very low. In those cases blocking or reporting as spam was enough. If such a system can be created, it might be great."* Some participants also point out that an abuse detection system for emails can be particularly beneficial for teenagers and young adults. P4 stated, *"Such a warning system may be very useful to teenagers and young adults because they might open abusive emails from curiosity."*

All of the interview participants and the majority of the survey participants (92%) expressed the need for an efficient abuse detection system for online platforms other than emailing platforms. However, only a few participants had previous experience of using abuse detection systems in emailing or any other platform. 5 out of 15 interview participants and 14% of the survey participants mentioned having such experience. This contradiction between demand and actual usage indicates several shortcomings of existing abuse detection systems from the view of the participants. Of those with experience of using an abuse detection system, most participants expressed their dissatisfaction over the inefficiency of current systems. Participants mentioned that often implicit abuses are subtle enough to remain undetected by detection systems. Moreover, the relationship between two people is an important determinant to whether or not something will be considered abusive. P7 explains, *"Language analysis is a big issue. Abuse is nowadays really subtle, so it may be hard to detect. Moreover, depending on the relationship between 2 persons, something can be abusive or not."* Some participants also pointed out that current systems are not advanced enough to detect abuse across multiple languages and cultures.

Participants also mentioned how most abuse detection systems provide reporting or blocking facilities but fail to prevent the abuse from happening in the first place. From the perspective of victims, even though current systems give control to them, they fail to protect them from the trauma inflicted. P2 explains, *"In daily life, I sometimes get abusive messages in Facebook Messenger and on Instagram. I have to actively delete those messages. But the problem is that* *to delete it, I have to read it and it's not a fun experience."* Participants further expressed that current abuse detection mechanisms make them feel less in control of the detection of the abusive content they receive. P11 said *"I think it would be better if the control existed in the hands of the user."* 37.5% of the survey responses also admitted feeling limited control in detecting abusive contents.

4 interview participants expressed frustration over the delay caused by existing systems in detecting and taking action against the abusive content and the abuser. P10 stated, *"There is no guarantee that an abusive content will be moderated for sure. Actions are delayed often."* Among the survey participants 13% participants identified delay as a shortcoming of existing abuse detection systems.

From our interviews and survey, we find out the preferred type of abuse detection system for emailing platforms. From a total of 15 participants, 6 of them wanted an automated system, 1 of them wanted a human moderator-based system, 8 of them wanted a hybrid system incorporating both automated and human-moderator-based detection. From the survey, we found that among the 65 responses, around 41% want an automated system, 8% want a human moderator-based system, 51% want a hybrid system with human-moderator-based and automated detection incorporated, for the task of abusive behavior detection for emailing platforms. Therefore, in both studies, there was a clear preference for a hybrid system that incorporates both automated and human-moderator-based detection. A hybrid system can incorporate the positive aspects of both types and reduce the negative aspects. P1 said, *"An automated system might label harmless emails from friends as abusive. Again, for a human moderator-based system, if someone receives a lot of emails, manually going through them might not be possible. I think there should be a good balance. An automated system can shortlist and human moderation can filter further."*

4.4 Strengths of Existing Abuse Detection Systems, especially over Emailing Platforms

While discussing their preferred types of abuse detection system with a particular focus on emailing platforms, the participants explained the factors that influenced their preferences.

Many participants preferred an automated abuse detection system because it grants them more control than other human moderators, who may or may not relate to their definition of abusive behavior. Survey participants reported limited control over the prevention of abusive behavior as the major limitation of current abuse detection systems. Therefore, to a significant section of participants, having control over what content gets detected as abusive or not is very critical, which is identified as a key strength of an automated abuse detection system.

From our interviews, we find that most users prefer fast detection of abusive behaviors. For time-sensitive modes of communication such as emails, users prefer automated detection, as it can provide near-instantaneous detection by not depending on any human moderator. P12 said *"Automated systems are needed so that I can take action by myself, as soon as I receive something offensive."*

Participants reported having fewer privacy concerns for automated abuse detection systems compared to that with human-moderator-based ones, as shown in Figure 3. On the other hand, the main strength identified for human-moderator-based detection is the involvement of human-in-the-loop for accurate detection of abusive behavior. Many participants were concerned that an automated abuse detection system may not be efficient enough to accurately detect abusive behavior across multiple languages and understand cultural nuances. However, a human moderator can take better decisions in this regard. P3 mentioned why human moderator-based systems may be preferable, *"Automated systems may not be able to detect abuse if I change the text, manual moderation by moderators may be more effective."*

4.5 Challenges of Existing Abuse Detection Systems, especially over Emailing Platforms

We find out the limitations felt by the participants in the existing types of abuse detection systems, with a special focus on abusive behaviors over emails.

The major concern that participants expressed for automated abuse detection systems were that they feel their personal information might get stored for the training of machine-learning-based automated systems, making their confidential information vulnerable. P1 mentioned, *"I will not give permission to store my emails. Often there are personal or sensitive credentials. For example, I share my flight information, travel insurance, ticket information with my travel agency. I often share personal information with my family members."* Fear of information breach is a concern for automated abuse detection in other platforms as well. P1 expressed, *"An automated system might access my contact list for learning purposes, that might raise privacy concerns for those people as well. If from my account something like a spam mail goes out to them, it may cause a problem for them."* Participants also shared concerns about unwanted advertisements while using such a system. Another major concern raised in the case of automated detection is its inefficiency in accurately detecting abuse taking into language, the relation between the concerned people, and cultural factors into consideration.

On the other hand, while discussing the limitations of a human-moderator-based detection system, participants expressed their exasperation that whether the content will be labeled as abusive or not, is solely in the hands of the moderator. This factor contributes to their preference for an automated or hybrid detection system, which has the potential to provide some control to the hands of the users. As P12 pointed out, *"Content that seems abusive to me, may not seem abusive to a human moderator."*

Participants state several privacy concerns regarding the human-moderator-based abuse detection systems. In the interviews, participants state they feel their personal information is at higher risk of getting exposed when human moderators are included. P6 mentioned, *"My private emails may get exposed in this way, and professional information night gets revealed."* P10 raised privacy concerns even in case of using a combined system, *"What if through the moderators, my story gets revealed and I am exposed to society? Yes, moderators may be unknown, but what guarantee is there?"*

In case of considering privacy concerns, human-moderator-based abuse detection system got 72% responses of maximum level (5) of concern whereas automated abuse detection system got maximum of 30% responses equally in both level 3 and level 5. Moreover, in opting for a human-moderator-based abuse detection system, "Friend" get the highest level of comfort (3 out of 5 levels) among options such as family member, paid moderator, unpaid moderator, known moderator, and unknown moderator. Some participants also felt that the time-sensitive nature of emails makes human-moderator-based inappropriate.

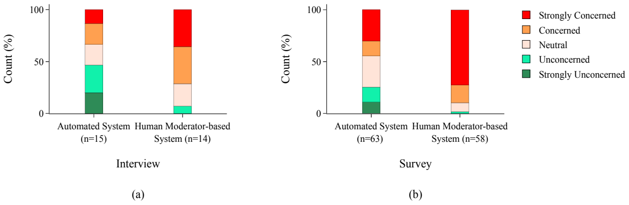

Figure 3: Comparison of participant's level of concern with privacy issues over using a human moderator-based and automated abuse detection system in emailing platforms (Response scale: 1=Strongly Unconcerned, 5=Strongly Concerned)

4.6 Expectations from an Abuse Detection System, especially over Emailing Platforms

Expected features in an abuse detection system for emailing platforms include showing the abusive emails in the notification window on being received in the inbox, options for contacting their trusted people, and exempting detection for chosen people.

Among the interview participants, 12 participants desired the feature to display abusive emails in the notification window after detection, 10 participants expected the option for contacting relatives and 11 participants expected the option for contacting friends. Figure 4 represents the comparison of expected features of the interview participants.

Many participants wanted to exempt their trusted contacts from detection. They expressed that depending on the relation between the people, content can seem abusive or not abusive, therefore, a mechanism to customize the target abuser pool should exist. 97% of the survey participants opted for being able to tune the strictness of the abuse detection system, depending on the relationship. P1 stated, *"If somehow my friends or trusted people can be excluded from the filtering step, it would be great."* P3 added, *"Some sort of two-way verification can be implemented so that I can confirm that an email detected as abusive is abusive or not from my perspective."* Incorporating information regarding the relationship between the concerned people to the machine-learning-based automated detection system in real-time was also suggested. Real-time warning for abusive content was also suggested. P7 suggested, *"While we are typing an email, an automated suggestion telling whether it is abusive or not can be great. Often we do not know what we are trying to say, so that type of real-time suggestion would be a great thing to integrate into an abuse detection system."*

Some participants desired options to contact and seek help from trusted people and law enforcement agencies on experiencing online abuse. Sending warning emails to the abuser was also suggested. A dashboard to view safe and unsafe emails as well as statistical data on experienced abusive content was suggested as well. From the survey, we discover that 72.3% of participants expect showing of an abusive email in the notification window after the detection, 46.2% expect contacting friends, 33.8% expect contacting relatives after detection of an abusive email among the 65 responses. Figure 4 represents the comparison of these top three suggested features among the expected features reported in the user studies.

The majority of participants in both interviews and survey opted for a hybrid system, therefore, expectations were based on the desire to have a system that incorporates the strengths of both types of abuse detection systems. Thus, a hybrid system was preferred that combines the fewer privacy concerns of automated systems and efficient detection that comes with human-in-the-loop involvement of human-moderator-based systems.

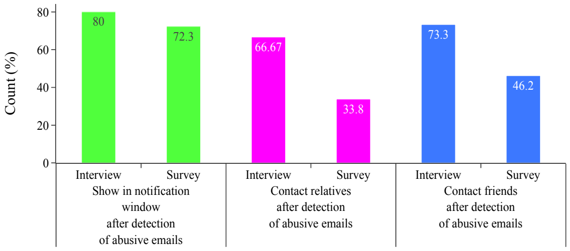

Figure 4: Comparison of participant's top three expected features in an abuse detection system in emailing platforms over interview (n=15) and survey (n=65)

5 DISCUSSION

In the previous sections, we have presented how online abuse affects the users and how they cope with the situation. We have shown how they feel about existing abuse detection systems, their concerns, and their expectations to prevent those concerns. We have further focused on the role of how an abuse detection system incorporating users' expectations can help them combat abuse over emails. In this section, we unravel a deeper understanding of users' interaction with online abuse, and existing abuse detection systems, particularly over emailing platforms.

5.1 Effects of Online Abuse, Coping Strategies, and Contradiction between Desire for an Abuse Detection System for Emails and Frequency of Abuse Occurring over Emails

Overall, the participants in our study reported experiencing online abuse often and understood their online risks fairly comprehensively. However, our participants strategically employed a range of coping mechanisms such as relying on family, relatives, friends, blocking, restricting the sources, seeking help from law enforcement agencies, etc. We echo that such expected roles of the near ones as well as the agencies need to be embraced in any interventions [29, 39]. We also find that many participants are not satisfied with the coping practices they follow in their everyday life on online platforms, as found from our interviews.

However, our study found that 40% of interview participants and only 11% of survey participants mentioned experiencing abuse over emails. This inconsistency among interview and survey participants can be attributed to prior studies showing that victims of online abuse in South-East Asia are uncomfortable sharing their experiences with anyone other than with their friends and family [41]. Since survey participants were unknown to the authors, the lower percentage of participants reporting abuse supports that the victims of online abuse in this region were indeed reluctant to share sensitive details of their experience outside their inner circle. Mostly recruited through personal networks, interview participants were more likely to mention their experiences with abuse, which is justified by the higher percentage reporting abuse.

Moreover, even though the extent of experiencing abuse over emails is less compared to that over other online platforms, expectations on abuse detection systems equally apply for the emailing platforms and other online platforms. Our study showed that email is an extensively used mode of communication in Bangladesh, yet, online abuse is less frequent over emails. However, that does not stop our participants from anticipating an abuse detection system on this platform. This is reflected through the responses of the users related to users' desirability of an abuse detection system for emails and shortcomings of existing solutions. Our results show that users are eager to use abuse detection systems on other platforms as well.

Another important point revealed in our study is that even though teens and young adults are groups mostly impacted by online abuse [25], desire for abuse detection systems spans to all ages. This becomes evident as our participants of diversified ages expressed their desires for abuse detection systems. As users, they want to get online abuse detected and feel the need for technological interventions in this regard. Thus, the desire for abuse detection systems goes beyond any specific age range.

In interviews, an equal ratio exists among males and females reporting abuse other than in emailing platforms, and 33.33% are female and 66.67% are male among those reporting abuse over emails. In the survey, 39.47% are female and 60.53% are male among those reporting abuse other than in emailing platforms, and 57.14% are female and 42.86% are male among those reporting abuse over emails. Therefore, we find no meaningful gender-based patterns in the responses from both user studies.

5.2 Efficiency and Privacy Dilemma: Conflict between Expectation and Reality

Many participants expressed that what may seem abusive to them may not seem to be so to the human moderator. Even then, the human moderators remain more efficient to detect abuse compared to the automated systems because the diversified nature of abuse is yet to be detected automatically with high efficiency [27].

However, on the other side of the coin, human-moderator-based systems raise privacy concerns as well as trust issues and give less control to the users. Automated systems are not free from privacy concerns as well where the storage of users' data can create serious privacy breaches. However, automated systems can introduce more control to the users with real-time preventive mechanism as we showed in our study. We also showed that users look for real-time preventive mechanisms with fewer privacy concerns over emailing platforms (Section 4.5). This expectation arises owing to the time-sensitive nature of the online platform, where introducing human moderators might introduce a time delay that does not happen in case of automated systems. As a result, these phenomena of having simultaneous preferences over efficiency and privacy by the users while these performance metrics remain a conflicting one in reality present a unavoidable dilemma.

5.3 Design Choices over Automated, Human-moderator-based, and Hybrid Abuse Detection Systems

From this study, we find that almost all participants face online abuse and demand for abuse detection systems, however, they are not satisfied over the existing abuse detection systems. Users feel having less control to prevent online abuse through existing mechanisms, feel privacy concerns, trust issues, and less comfort regarding human-moderator-based abuse detection systems. However, they also feel that accurate detection of abuse needs human moderators, which is perhaps not possible in a fully automated system.

We call for HCI research to take a deeper look into this issue, HCI research can focus on designing tools and techniques, which will protect the users from the effects associated with online abuse engendering fewer privacy concerns. From our user study, we also see users ask to own more control while preventing online abuse. Therefore, we can propose some design components to provide them have more control. A feature to having real-time connecting with trusted contacts (friends, family, etc.), law enforcement agencies, etc., in case of facing online abuse can provide them more control as presented in Section 4.6.

Besides, we can introduce the expected feature of having user control by enabling a threshold for filtering abusive contents and enabling blocking contacts of abusers. In our study, many participants wanted customization capability such as flagging some words, which they believe abusive. Participants felt seeing the abusiveness ratings of the abusive text can also make them feel more in control since they can know for sure those texts were abusive for which they can take action against the abusers. Considering the social and mental conditions of the users, the design of such tools should incorporate more control to the users and generate fewer privacy concerns. However, one thing becomes evident from all these expectations and interpretations that perhaps neither automated nor human-moderator-based abuse detection systems can entertain the expectations resulting in an unavoidable need for a hybrid system. The hybrid system will need to realize a delicate balance between all the expectations at the best possible levels.

6 LIMITATIONS OF THE WORK

We faced some limitations in ensuring the diversity in the participants of the interviews. Most of the interview and survey participants are mostly from researchers' primary and secondary networks since they were recruited through snowball sampling. Therefore, our work is not free from participation bias and selection bias. Thus, arguments and opinions driven from the interviews and survey may not represent the collective view of the people of the whole nation. Despite these limitations, the findings of our study will be useful for technology design in the context of online abuse.

7 SCOPE OF FURTHER RESEARCH

In the future, we plan to expand the scale of our research by incorporating more participants in our user studies. Since studies show that victims of online abuse in South-East Asia are uncomfortable in sharing their experiences with abuse other than with friends and family [41], most of our interview participants have been recruited through personal networks. In the future, we intend to include victims of online abuse going beyond our known circles and further research ways to include them in user studies by studying the barriers victims of this region face in sharing sensitive details of their experiences outside their inner circle. We also intend to include participants from other countries and explore the differences in their perceptions. Moreover, based on the findings from the interviews and the survey, our next step is to design and implement an abuse detection and prevention system for emailing platforms. To confirm our findings from this study, we then plan to evaluate the system through usability testing. The findings of the evaluation phase will provide valuable insights into the design of an abuse detection and prevention system.

8 CONCLUSION

Online abuse has become very prevalent with the development of online platforms. Although technological strategies exist, these are sometimes ineffective and unsatisfactory. Users face difficulties through existing online abuse detection systems to prevent online abuse. Though users face a lesser frequency of abuse over emails, they feel a need for abuse detection systems for this platform. Privacy issues with human moderators, less control to users introduce challenges for the users. Through extensive literature study, we confirm that creating technology-based solutions to cater to their concerns is, therefore, a critical need. We conduct an interview study challenges faced by users in existing online abuse detection systems especially over emailing platforms. Subsequently, we conduct an online survey to find out the familiarity of existing abuse detection systems and the challenges they identify in the existing solutions especially over emailing platforms. While several abuse detection systems with some preventive mechanisms are prevalent worldwide, expectations from users' perspectives are still to be met. We present several insightful findings from our rigorous user studies and discuss how new technological interventions may be designed to address the concerns of users.

REFERENCES

[1] 2019. Tune (experimental). https://chrome.google.com/webstore/detail/tune-experimental/gdfknffdmmjakmlikbpdngpcpbbfhbnp?hl=en/.

[2] Eileen Alexy, Ann Burgess, Timothy Baker, and Shirley Smoyak. 2005. Perceptions of Cyberstalking Among College Students. *Brief Treatment and Crisis Intervention* 5 (08 2005). https://doi.org/10.1093/brief-treatment/mhi020

[3] Daniel Etcovitch Andrew Arsht. 2018. The Human Cost of Online Content Moderation. https://jolt.law.harvard.edu/digest/the-human-cost-of-online-content-moderation/.

[4] Jennifer Beckett. 2018. We need to talk about the mental health of content moderators. https://theconversation.com/we-need-to-talk-about-the-mental-health-of-content-moderators-103830.

[5] Lindsay Blackwell, Jill Dimond, Sarita Schoenebeck, and Cliff Lampe. 2017. Classification and Its Consequences for Online Harassment: Design Insights from HeartMob. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW, Article 24 (Dec. 2017), 19 pages. https://doi.org/10.1145/3134659

[6] Michael L. Bourke and Sarah W. Craun. 2014. Secondary Traumatic Stress Among Internet Crimes Against Children Task Force Personnel: Impact, Risk Factors, and Coping Strategies. *Sexual Abuse* 26, 6 (2014), 586–609. https://doi.org/10.1177/1079063213509411 arXiv:https://doi.org/10.1177/1079063213509411 PMID: 24259539.

[7] Richard Boyatzis. 1998. Transforming qualitative information: Thematic analysis code development. sage, thousand oaks. (01 1998).

[8] Eshwar Chandrasekharan, Chaitrali Gandhi, Matthew Wortley Mustelier, and Eric Gilbert. 2019. Crossmod: A cross-community learning-based system to assist reddit moderators. *Proceedings of the ACM on human-computer interaction* 3, CSCW (2019), 1–30.

[9] Eshwar Chandrasekharan, Mattia Samory, Anirudh Srinivasan, and Eric Gilbert. 2017. The bag of communities: Identifying abusive behavior online with preexisting internet data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 3175–3187.

[10] J. Clement. 2020. Number of e-mail users worldwide 2017-2024. https://www.statista.com/statistics/255080/number-of-e-mail-users-worldwide/#statisticContainer.

[11] M. Dadvar and Franciska M.G. de Jong. 2012. Cyberbullying Detection; A Step Toward a Safer Internet Yard. In *Proceedings of the 21st International World Wide Web Conference, WWW 2012 - PhD-Symposium*. Association for Computing Machinery (ACM), United States, 121–125. https://doi.org/10.1145/2187980.2187995 null ; Conference date: 16-04-2012 Through 20-04-2012.

[12] Maeve Duggan. 2014. Online Harassment. https://www.pewresearch.org/internet/2007/06/27/cyberbullying/.

[13] Maeve Duggan. 2017. Online Harassment. https://www.pewresearch.org/internet/2017/07/11/online-harassment-2017/.

[14] Jeanne Whalen Elizabeth Dwoskin and Regine Cabato. 2019. Content moderators at YouTube, Facebook and Twitter see the worst of the web — and suffer silently. https://www.washingtonpost.com/technology/2019/07/25/social-media-companies-are-outsourcing-their-dirty-work-philippines-generation-workers-is-paying-price/.

[15] Facebook. 2018. Publishing Our Internal Enforcement Guidelines and Expanding Our Appeals Process. https://about.fb.com/news/2018/04/comprehensive-community-standards/.

[16] Facebook. 2020. Blocks. https://www.facebook.com/help/174623239336651/.

[17] FaceBook. 2020. Facebook Terms and Policies. https://www.facebook.com/communitystandards/.

[18] Tarleton Gillespie. 2018. *Custodians of the internet: Platforms, content moderation, and the hidden decisions that shape social media.* 1–288 pages.

[19] Kayla Hendricks, Pitso Tsibolane, and Jean-Paul Van Belle. 2020. Cyber-Harassment Victimization Among South African LGBTQIA+ Youth. In *Conference on e-Business, e-Services and e-Society*. Springer, 135–146.

[20] Reddit Inc. 2020. Moderator Guidelines for Healthy Communities. https://www.redditinc.com/policies/moderator-guidelines/.

[21] Shagun Jhaver, Darren Scott Appling, Eric Gilbert, and Amy Bruckman. 2019. "Did You Suspect the Post Would Be Removed?": Understanding User Reactions to Content Removals on Reddit. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 192 (Nov. 2019), 33 pages. https://doi.org/10.1145/3359294

[22] Rachael Krishna. 2018. Tumblr Launched An Algorithm To Flag Porn And So Far It's Just Caused Chaos. https://www.buzzfeednews.com/article/krishrach/tumblr-porn-algorithm-ban/.

[23] Joanne Kuzma. 2013. Empirical study of cyber harassment among social networks. *International Journal of Technology and Human Interaction (IJTHI)* 9, 2 (2013), 53–65.

[24] Cliff Lampe and Paul Resnick. 2004. Slash (dot) and burn: distributed moderation in a large online conversation space. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 543–550.

[25] AMANDA LENHART. 2007. Cyberbullying. https://www.pewresearch.org/internet/2007/06/27/cyberbullying/.

[26] Amanda Lenhart, Michele Ybarra, Kathryn Zickuhr, and Myeshia Price-Feeney. 2016. *Online harassment, digital abuse, and cyberstalking in America*. Data and Society Research Institute.

[27] Emma J Llansó. 2020. No amount of "AI" in content moderation will solve filtering's prior-restraint problem. *Big Data & Society* 7, 1 (2020), 2053951720920686.

[28] Kaitlin Mahar, Amy X Zhang, and David Karger. 2018. Squadbox: A tool to combat email harassment using friendsourced moderation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–13.

[29] Niveditha Menon. 2008. Domestic violence in India: Identifying types of control and coping mechanisms in violent relationships. (2008).

[30] Casey Newton. 2019. THE TERROR QUEUE. https://www.theverge.com/2019/12/16/21021005/google-youtube-moderators-ptsd-accenture-violent-disturbing-content-interviews-video/.

[31] Casey Newton. 2019. THE TRAUMA FLOOR. https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona/.

[32] Casey Newton Nitasha Tiku. 2015. Twitter CEO: 'We suck at dealing with abuse. https://www.theverge.com/2015/2/4/7982099/twitter-ceo-sent-memo-taking-personal-responsibility-for-the.

[33] Justin Patchin and Sameer Hinduja. 2010. Cyberbullying and self-esteem. *Health* 80 (01 2010), 616–623.

[34] Jessica A Pater, Moon K Kim, Elizabeth D Mynatt, and Casey Fiesler. 2016. Characterizations of online harassment: Comparing policies across social media platforms. In *Proceedings of the 19th international conference on supporting group work*. 369–374.

[35] Karen L Paullet, Daniel R Rota, and Thomas T Swan. 2009. Cyberstalking: An exploratory study of students at a mid-Atlantic university. *Issues in Information*

Systems 10, 2 (2009), 640–649.

[36] ROBERT PECK. 2019. The Punishing Ecstasy of Being a Reddit Moderator. https://www.wired.com/story/the-punishing-ecstasy-of-being-a-reddit-moderator/.

[37] Benjamin Plakett. 2018. Unpaid and abused: Moderators speak out against Reddi. https://www.engadget.com/2018-08-31-reddit-moderators-speak-out.html?guccounter=1/.

[38] Twitter Public Policy. 2018. Evolving our Twitter Transparency Report: expanded data and insights. https://blog.twitter.com/official/en_us/topics/company/2018/evolving-our-twitter-transparency-report.html.

[39] Fauzia Rabbani, F Qureshi, and Narjis Rizvi. 2008. Perspectives on domestic violence: case study from Karachi, Pakistan. *EMHJ-Eastern Mediterranean Health Journal, 14 (2), 415-426, 2008* (2008).

[40] Sarah T Roberts. 2014. *Behind the screen: The hidden digital labor of commercial content moderation.* Ph.D. Dissertation. University of Illinois at Urbana-Champaign.

[41] Nithya Sambasivan, Amna Batool, Nova Ahmed, Tara Matthews, Kurt Thomas, Laura Sanely Gaytán-Lugo, David Nemer, Elie Bursztein, Elizabeth Churchill, and Sunny Consolvo. 2019. " They Don't Leave Us Alone Anywhere We Go" Gender and Digital Abuse in South Asia. In *proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14.

[42] Ari Schlesinger, Eshwar Chandrasekharan, Christina A. Masden, Amy S. Bruckman, W. Keith Edwards, and Rebecca E. Grinter. 2017. *Situated Anonymity: Impacts of Anonymity, Ephemerality, and Hyper-Locality on Social Media*. Association for Computing Machinery, New York, NY, USA, 6912–6924. https://doi.org/10.1145/3025453.3025682

[43] Inc Twitter. 2020. Report abusive behavio. https://help.twitter.com/en/safety-and-security/report-abusive-behavior.

[44] Inc Twitter. 2020. The Twitter Rules. https://help.twitter.com/en/rules-and-policies/twitter-rules/.

[45] The Verge. 2016. THE SECRET RULES OF THE INTERNET. https://www.theverge.com/2016/4/13/11387934/internet-moderator-history-youtube-facebook-reddit-censorship-free-speech/.

[46] Janis Wolak, Kimberly J Mitchell, and David Finkelhor. 2007. Does online harassment constitute bullying? An exploration of online harassment by known peers and online-only contacts. *Journal of adolescent health* 41, 6 (2007), S51–S58.

[47] Queenie Wong. 2019. Facebook content moderation is an ugly business. Here's who does it. https://www.cnet.com/news/facebook-content-moderation-is-an-ugly-business-heres-who-does-it/.

[48] YouTube. 2020. Community Guidelines. https://www.youtube.com/howyoutubeworks/policies/community-guidelines/.