Proceedings of the Eighteenth International AAAI Conference on Web and Social Media (ICWSM 2024)

Assessing the Impact of Online Harassment on Youth Mental Health in Private Networked Spaces

Seunghyun Kim[1], **Afasneh Razi**[2], **Ashwaq Alsoubai**[3],
Pamela J. Wisniewski[3], **Munmun De Choudhury**[1]
[1]Georgia Institue of Technology,
[2]Drexel University,
[3]Vanderbilt University
seunghyun.kim@gatech.edu, afsaneh.razi@drexel.edu, ashwaq.alsoubai@vanderbilt.edu,
pamela.wisniewski@vanderbilt.edu, munmund@gatech.edu

Abstract

Online harassment negatively impacts mental health, with victims expressing increased concerns such as depression, anxiety, and even increased risk of suicide, especially among youth and young adults. Yet, research has mainly focused on building automated systems to detect harassment incidents based on publicly available social media trace data, overlooking the impact of these negative events on the victims, especially in private channels of communication. Looking to close this gap, we examine a large dataset of private message conversations from Instagram shared and annotated by youth aged 13-21. We apply trained classifiers from online mental health to analyze the impact of online harassment on indicators pertinent to mental health expressions. Through a robust causal inference design involving a difference-in-differences analysis, we show that harassment results in greater expression of mental health concerns in victims up to 14 days following the incidents, while controlling for time, seasonality, and topic of conversation. Our study provides new benchmarks to quantify how victims perceive online harassment in the immediate aftermath of when it occurs. We make social justice-centered design recommendations to support harassment victims in private networked spaces. *We caution that some of the paper's content could be triggering to readers.*

Introduction

Online harassment–a form of abusive treatment online, which encompasses cyberbullying, hate speech, and threats of violence–is a prevailing problem that causes unfavorable experiences for many users on social media platforms (Lewis, Zamith, and Coddington 2020). Negative experiences from online harassment can have long-lasting consequences, such as psychological distress, depression, and increased risk of suicide, especially among youth and young adults (Brody 2021; Aponte and Richards 2013). Consequently, there has been extensive research on developing effective automated detection systems for online harassment (Rosa et al. 2019). Natural language processing (NLP) and machine learning (ML) have been utilized on popular social media platforms such as Twitter and Reddit (Chatzakou et al. 2017; Almerekhi, Jansen, and Kwak 2020). The effects of online harassment on an individual, both short-term and long-term, has also been explored in related fields such as psychology and social science, primarily via the use of retrospective self-reports and interviews (Rigby 2003).

The vast, complex, and unstructured nature of social media data, makes the manual monitoring for such online risk impractical (Kumar and Sachdeva 2019). Consequently, the urgency to counteract these adverse activities and its enduring harms has propelled research towards the development of automated detection systems. This shift towards technological solutions aims to effectively identify incidents of cyberbullying, highlighting the intersection of computational methods and social welfare in contemporary research endeavors. Prior literature has developed automated systems to detect online harassment (Kim et al. 2021a; Rosa et al. 2019); yet, the use of computational approaches to assess the mental health impacts resulting from online harassment is nascent. *How* individuals are impacted from these online harassment incidents in comparison to those that do not experience online harassment has yet to be systematically explored via their digital trace data. Moreover, due to the sensitive, traumatizing, and stigmatizing nature of the experience, victims' recollection of past events and impacts may be difficult to gather accurately and present ethical complexities. Reliance on large-scale naturalistic data where both harassment and mental health expressions are quantifiable and observable could close this gap. So far, such empirical investigations are limited due to such constraints.

Scholars have noted how victims of traumatic events seek support from others both directly and indirectly on the web, disclosing their traumatic events anonymously (Andalibi et al. 2018). When public, such channels tend to provide a window for the victims to receive direct support to help overcome their sufferings. When private, these networked spaces may act as a cathartic relief by enabling sharing sensitive information without having to worry about the risk of context collapse or being identified (Younas, Naseem, and Mustafa 2020). Yet, few researchers have studied how people express their mental state in their online communications in *private* channels and how these expressions could have causal associations with experiences of harassment on the same channels (Adams et al. 2022; Naslund et al. 2020). To that end, we posit that there are notable differences between people

who receive online harassment and those who do not, in regards to how they express their mental statuses in private networked spaces; however, these differences have yet to be systematically studied. Accordingly, we pose the following high-level research questions:

RQ1: *How can we assess differences in mental health disclosures made in private social media conversations by individuals who reported harassment versus those who did not?*

RQ2: *How does an online harassment incident impact subsequent mental health expressions made in private contexts?*

To answer these questions, we utilized a rich dataset of private conversations donated by young (ages 13-21) Instagram users, consisting of over 1.4 million messages annotated by participants themselves for online harassment incidents. We divided youth into two groups based on whether they received any online harassment messages or not within their private message conversations. We adopted and subsequently validated a transfer learning approach, based on a machine learning model trained on Reddit data, to assess and quantify the levels of mental health expressions in private messages, thereafter contrasting the mental health of those victimized by online harassment and those not. Then, adopting a causal inference framework where the outcome of interest was the degree of mental health expression, we created a *treatment* group of those Instagram users who were harassed, and two controls. The first was a *within-subjects control* consisting of the same users from a time period distant from when they were harassed, and second was a matched *between-subjects control* of other users who were never harassed but with similar topical usage of Instagram during a similar seasonal period. Upon performing a difference-in-differences analysis (Angrist and Pischke 2009) along with modeling two types of counterfactuals, we observed that compared to the two control groups, the treatment group expressed statistically significantly aggravated negative mental health attributes compared to the *within-subjects control* and *between-subjects control*, in a period of 7 and 14 days following harassment. We outline our key contributions:

- We assess the way people depict their mental health in private messages by using transfer learning and validating it with self-reported data from standardized psychometric instruments. We showcase face validity of this approach by demonstrating harassed individuals experiencing greater mental health concerns than others.
- We provide insights into the relative impact of online harassment on the way people express their mental health in private conversations, using robust causal inference techniques. Harassment results in worsened mental health in victims for a period upto 2 weeks compared to the same duration before – a pattern not observable in the controls without experience of harassment.

Broader Perspective, Ethics, and Competing Interests. Our findings shine light on the deleterious impacts of online harassment on the mental health of victims, providing a robust causal approach applied to a unique and rich dataset of private online interactions of youth. The study's findings have important implications for the stakeholders of automated detection systems designed to utilize large-scale web data to identify incidences of problematic behaviors to also examine and potentially intervene to mitigate ensuing negative psychological outcomes. That said, such interventions would need to adopt a victim-centered approach, in order to ensure they are ethical and do not result in unintended harmful consequences. This study was approved by the Institutional Review Board (IRB) of the authors' institution. We took measures to protect the privacy of the participants of the study by removing any personally identifiable information and paraphrasing any reported text.

Related Work

Online Harassment Research

Numerous scholars have utilized mixed methods, such as qualitative interviews or quantitative surveys, to examine users' circumstances leading to and experiences of online harassment. Techniques such as concurrent note-taking, reflective observation, and thematic analysis have been used to gain a more comprehensive understanding of the context and dynamics of online harassment (Rahman 2020; Hafford-Letchfield, Toze, and Westwood 2022; Harasgama and Jayamaha 2023). Researchers have also utilized machine learning and natural language processing approaches to identify and categorize online harassment. A majority of these computational approaches used supervised detection methods that incorporated a range of textual features such as lexicon, sentiment, and word embeddings (Chatzakou et al. 2017; Almerekhi, Jansen, and Kwak 2020).

Each of these approaches have limitations; qualitative methods have limited capacity to incorporate broader contextual factors surrounding the studied phenomenon (Rahman 2020). The approach of solely using surveys where the responses typically comprise recollections of past experiences are vulnerable to potential recall bias or observer-expectancy effect (De Choudhury 2013). Such limitation becomes particularly relevant when it comes to understanding the lived experiences of online harassment victims on social media. Existing research on detecting online risk in social media, on the other hand, has mainly relied on publicly available datasets to create multiple benchmarks that are utilized across a wide body of research aimed to improve detection models or intervention designs (Kim et al. 2021a). While this approach allows researchers to avoid ethical or legal challenges that could come from direct interaction with users, scholars have noted the unique differences between public and private networked spaces and how they might influence the way users communicate and share information (Kim et al. 2021b). Moreover, there has been a primary focus on identifying incidents, rather than exploring the subsequent interactions between individuals involved in these events. As a result, the impact of online harassment on victims in these (private) networked spaces is less understood (Yao, Chelmis, and Zois 2018; Dinakar, Reichart, and Lieberman 2011). Our study addresses these gaps in both approaches by examining a large-scale dataset of private messages donated and annotated by youth and assessing the impact of harassment across different time periods through a causal analysis.

Assessing the Impact of Online Harassment

Researchers have questioned and explored the influence of online harassment, identifying the negative impacts of online harassment across multiple platforms as well as victims (Stevens, Nurse, and Arief 2021; Cañas et al. 2020). There have been longitudinal studies that examine the causal relationship between online harassment and mental health (Ståhl and Dennhag 2021; Hemphill, Kotevski, and Heerde 2015; McHugh et al. 2017); however, most of them have often involved self-reports or surveys to assess the impact on the victims, which, as mentioned above, are prone to memory bias and observer-expectancy effect. Stevens, Nurse, and Arief (2021), through the systematic review on the association between online harassment and mental health, advocated the importance of longitudinal research designs to explore casual relationships.

Causal investigations are valuable because, for instance, several longitudinal studies have shown individuals with mental health struggles are known to be already vulnerable to online harassment (Arseneault, Bowes, and Shakoor 2010) and thus the extent to which such negative incidents may exacerbate or trigger new mental health symptoms can have widespread implications for how we care for victims. Furthermore, there is a need to explore the causal relationship within social media trace data, which offers a more objective and naturalistic form of observation compared to self-report data (De Choudhury 2013). By identifying causal patterns in social media trace data, we can gain a deeper understanding of the unfolding negative mental health impacts subsequent to online harassment incidents. Causality between online harassment and mental health, could only be assessed through multiple measurements over time (VanderWeele, Jackson, and Li 2016) Consistent with the goals of this paper, examining the causal relationship within the private networked space, therefore, would not only provide valuable insights into the association between mental health and online harassment, but also quantify the directionality of the relationship. Since we analyze individual-contributed and individual-annotated longitudinal data accounting for various confounds, it helps us overcome limitations surfacing from population-level snapshot-based understandings of the association between harassment and mental wellbeing.

Data Collection

Private Instagram Dataset

For this study, we analyzed a rich dataset of Instagram private message conversations gathered by Razi et al. (2022). To collect this data, the authors built a secured web-based system for participants to donate their Instagram archives and annotate their private conversations for unsafe messages. Participants were between 13 and 21 years old, based in the United States, and spoke English to be eligible. They also had at least 15 private conversations in Instagram and at least two of them were made them feel unsafe or uncomfortable. Informed consent was required by participants who were older than 18 and parents of participants who were under 18 years old while informed assent was required by teens. Once a participant completed the study and uploaded their data, multiple verification checks were implemented to assure the validity of the submitted data. Eligibility criteria such as whether there were at least two unsafe conversations, the time spent to complete the study, and if the conversations were between real people were applied. As a result, the dataset resulted in $n = 80$ verified participants with 11,267 conversations comprising 1,429,189 messages.

Victim Annotations of Harassment

The data of the 80 participants, from Razi et al. (2022) included participants' annotations of their own private messages and conversations in their Instagram archives. In particular, these authors invited participants to review their messages for self-assessment on various risk types, drawing upon Instagram's risk categories for reporting [1]. Specific categories included but were not limited to, Nudity/porn, Sexual messages/sexting or solicitations, Harassment, Hate speech, Violence/threat of violence, Sale or promotion of illegal activity, Self-injury, or Other. For the purpose of this study, we filtered the following risk types to include messages that were flagged as *Harassment*, *Hate speech*, and *Violence/Threat of violence*. We collectively refer these three categories as Harassment throughout this paper.

Self-Reported Data on Psychometric Instruments

In addition to the donation of their Instagram data, each participant was prompted to provide responses to a set of standardized questionnaires that described their interaction behavior with social media platforms as well as their mental health status. The following particular psychometric instruments were used, a set widely used in mental health and psychiatry research (Murphy et al. 2021): Short Warwick-Edinburgh Mental Wellbeing Scale (SWEMWBS), Inventory of Statements About Self-injury (ISAS), and Patient Health Questionnaire-9 (PHQ-9). The answers to each question in the survey was based on a 5-point Likert scale.

Measuring Mental Health in Private Messages

Model Training on Reddit Mental Health Data

To understand the way people with and without an experience of harassment depict their mental health in private conversations (RQ1), we first built a series of mental health classifiers from Reddit, following Saha et al. (2019); Saha and De Choudhury (2017). As we were interested not in the diagnosis of mental illnesses but rather the way people describe their mental health, a platform where people collectively talk about a mental health, such as Reddit, was suitable for the study. Moreover, since we intended to understand individual-level mental health expressions, ground truth was available in the form of participant responses to the above-noted psychometric surveys, rather than at the message or conversation level. This further justified a transfer learning approach of utilizing post-level mental health inferences from another platform.

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

Again per Saha et al. (2019), we first selected a total of seven subreddits that are related to mental health (MH subreddits): *r/anxiety, r/depression, r/psychosis, r/schizophrenia, r/selfharm, r/stress, r/SuicideWatch*. Then we collected all posts (or submissions) from the subreddits posted from January 1st, 2017 to January 1st, 2022 using the PushShift API, amounting to 913,485 posts in all. With these posts from MH subreddits as the positive class, we collected posts from the same time period from various subreddits that were on the main Reddit homepage to make a negative class, totaling 6,463,400 posts: *r/politics,r/AskReddit, r/nottheonion, r/aww, r/movies, r/IAmA, r/stocks, r/interestingasfuck, r/movies, r/coolguides*. We created a sampled subset (of 730,786 posts) for each of the positive subreddits in order to train binary classifiers. Adopting a transfer learning approach like Saha et al. (2019), the classifiers were trained with Linear Support Vector Machines (SVM) and used Linguistic Inquiry and Word Count (LIWC) as features, following prior studies that examined the relationship between online harassment and mental health in text messages (Liu et al. 2022; Stamatis et al. 2022). Each model used a *k*-fold ($k = 10$) cross-validation for hyperparameter tuning. Models were tested on a held out dataset comprising 91,350 positive and 91,349 negative examples.

Applying Transfer Learning to Instagram Data

We used a transfer learning approach to label the mental health expression in private Instagram conversations of users who experienced online harassment. To address the difference in message length between Instagram and Reddit, we grouped Instagram messages sent by the same participant by date and concatenated them. We obtained a mental health label for each message using the above described classifiers, and then summed the scores to assign a final score in the [0,7] range to each message, with larger scores indicating more mental health concerns. Finally, we aggregated the scores over all messages to obtain a single average mental health expression (MHE) score for each participant.

Classification and Transfer Learning Results

Table 1 summarizes the performance metrics of the classifcation models of mental health expression (MHE). All classifiers showed a test accuracy ranging from 0.80 and 0.93 with a F1-score range of 0.80 and 0.93.

Subreddit	Accuracy	F1-score	AUC
Anxiety	0.92	0.92	0.92
Depression	0.89	0.89	0.88
Psychosis	0.82	0.82	0.82
Schizophrenia	0.82	0.82	0.82
Selfharm	0.93	0.93	0.93
Stress	0.80	0.80	0.80
SuicideWatch	0.90	0.90	0.90
Mean	**0.87**	**0.87**	**0.87**

Table 1: MHE classification results in Reddit data.

Despite the strong performance of these classifiers on Reddit data, whether they transfer accurately to private networked spaces (Instagram private messages) necessitates further investigation. This is especially a pertinent question because Instagram is a platform with different features and affordances from Reddit, the most notable of which is that the former is private and the latter public. Thus the transferability of the mental health classifiers needs to be validated. Accordingly, we utilized a strategy to test the linguistic equivalence of the content shared on the two platforms, based on an approach developed by Saha et al. (2017). We first obtained the top 500 *n*-grams ($n = 1$) of each subreddit and the Instagram dataset. Using each *n*-gram's GloVe embeddings, we obtained a 200-dimensional vector space for each of these *n*-grams and then computed the cosine similarity between the *n*-gram embeddings of each subreddit and the Instagram data. Through this we were able to observe that there were high cosine similarities of the Instagram message content with that from each of the subreddits (avg. *0.99*), indicating linguistic equivalence.

While this demonstrates robustness in classifier transferability to (private) Instagram from (public) Reddit, additional validation is warranted to ensure that what is being classified in the Instagram private messages, is indeed indicative of mental health concerns. As noted above, we used the Reddit classifiers to automatically label each private message of each participant on Instagram, summing it to obtain a [0,7] range MHE score. Then, pursuing a concurrent validity testing approach, we utilized participants' responses on the standardized questionnaires on mental health: SWEMWBS, ISAS, and PHQ-9 (ref. Data Collection). These scales assess mental wellbeing, self-injury, and depression/suicide, thus being pertinent to the mental health classifiers we built on Reddit and transferred to Instagram. We aggregated an overall mental health score for each participant across the three scales by summing all responses. Together, the aforementioned two steps allowed us to obtain two scores per participant – one score coming from the transferred mental health classifiers applied to their private messages, and another score from their self-reported responses on the psychometric instruments.

On these pairwise scores, we conducted a Pearson's correlation test, which showed a statistically significant correlation ($corr = 0.29, p < 0.05$) between the self-reported mental health status of the participants and the computationally assessed MHE in Instagram messages. This correlation aligns with the findings from prior literature that studies the association between linguistic features and psychopathology from social media literature (Liu et al. 2022; Stamatis et al. 2022). Liu et al. (2022)'s study, which aimed to study the relationship between mental illness and online harassment, observed a close relationship between LIWC categories and depressive symptom severity ($r = 0.28, p < 0.001$). Furthermore, the effect size could be indicative of a more substantial population effect if the intervention, namely online harassment messages, is widely implemented (Matthay et al. 2021). Given the prevalence of online harassment among adolescents, the significance of the effect size should not be disregarded. In turn, this provides additional validity to the classifiers and our approach to mental health assessment expressed in private channels.

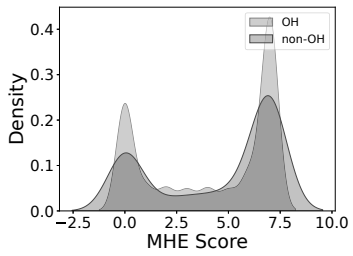


Figure 1: Kernel distribution graphs of the MHE scores for OH and non-OH participants.

Next, having validated this classifier, we present findings on how the levels of MHE differ across participants who did ("Online Harassment" or OH participants) and those who did not experience harassment ("Non-Online Harassment" or non-OH participants). Figure 1 gives a kernel density plot of MHE score over the two groups of participants. Both participant groups exhibit a bimodal distribution. However, notably, we observe a higher Fisher-Pearson skewness coefficient in the distribution of the OH participants compared to the non-OH participants (0.18, 29.17%), suggesting that the OH participants may be more prone to sharing messages with higher MHE scores. A Mann-Whitney test further verified this difference to be statistically significant between the two distributions ($U = 5218334.5, p < 0.05$).

Furthermore, we qualitatively looked into sample messages that received high/low MHE scores for the two participant groups. Since we grouped messages by the day they were sent, as shown in Table 2, across the two groups, messages that received a higher score (5 or higher) from the MHE classifiers showed increased usage of words indicative of negative affect, or loneliness. This aligns with findings in prior literature that demonstrates the close relationship between mental health, lowered self-esteem and self-efficacy, social isolation, and negative mood (Liberatore-Maguire et al. 2022). However, there are discernable differences between OH and non-OH participants' MHE, e.g., the high MHE messages from the OH participants show sadness and anger while those from the non-OH participants exhibit apologetic messages. These findings are consistent with prior research on linguistic differences associated with different mental states (Uban, Chulvi, and Rosso 2021).

These findings being correlational, RQ2 builds on these differences to assess if the experience of harassment itself could cause higher MHE scores in the OH participants.

Mental Health Impact of Harassment

Effectively answering RQ2 necessitates the use of causal methods to reduce biases associated with the observed effects (MHE) following harassment. A randomized control trial approach could address this issue; however, such a method's applicability in our work would be both impractical and unethical, due to the sensitive nature of both harassment and mental health experience. Therefore we adopt an observational study design – methods and frameworks which have been leveraged in prior social media studies (Saha et al. 2019). We utilize a causal inference methodology that employs matching. This technique aims to replicate the conditions of a randomized controlled trial by controlling for as many covariates as feasible (Imbens and Rubin 2015). The potential outcomes framework underpins this, which assesses whether a treatment T (harassment) causes an outcome (MHE) by comparing two potential outcomes: (1) $Y_i(T = 1)$ if T was applied, and (2) $Y_i(T = 0)$ if T was not applied. Under this framework, we thus constructed a single-blinded experimental setup where we divided the participants into *treatment* and *control* groups.

OH Participants	
High MHE	mind your own business like ???? I can't believe people are like this. [...] i don't get it. [...] I just think it's so nasal. that is my input. so if you're gonna attack me please hit me where it hurts.
	I just think about how many times I've cried alone in my room [...] I'm just upset that things turned out this way. [...] I'm having one of my sad days.
Zero MHE	Sad MC Moment. I know but its so tedious. Its be poggers. Do Hampen Parks by e-dubble please
	Good that you're good yo. Just checking on people yo. It's okay yo. How you doing
nonOH Participants	
High MHE	I'm going to leave now that you think I'm literally a sociopath. [...] I have issues. [...] twist as hard as the can and put me out of my misery.
	I'm sorry [...] it gives me a lot of anxiety, I don't think I would be good company
Zero MHE	[...] Hey! We're looking for brand ambassadors and models, send a message to our main account.
	Just request whatever [...] I can't do anything after the 26th cause imma be on vacation for 2 weeks.

Table 2: Examples of paraphrased private messages with high (≥ 5) and zero mental health expression scores of OH and non-OH participants.

Constructing Treatment and Control Groups

We divided the users into two sets–those who received messages that were flagged as harassment and those that never received such message. Out of the 80 participants of this study, a total of 74 had received at least one harassment message (OH participants) and 6 did not (non-OH participants). *Treatment group.* The OH participants constituted our *treatment* group. For each OH participant, we then created pre-OH and post-OH pairs of messages of the participant around each OH message that the participant received in relation to a given time window of $n = 7, 14, 21, 28$ days. We did not divide the messages by conversation and collectively created one set of pre-OH and post-OH messages for a given OH message that spanned all conversations in the $n = 7, 14, 21, 28$ day windows preceding and succeeding the OH message. Our rationale is as follows. While people could talk about vastly different subjects over different conversations, experiencing online harassment may trigger spillover effects, meaning, any ensuing mental health impact can surface not only in the same conversation with the perpetrator, but also in other temporally adjacent conversa-

tions with others. Moreover, we observed that in many cases, the user would retreat from the conversation right after the OH message, and thus to identify any mental health impact, considering other private conversations was imperative as it provided a comprehensive picture of the impact.

Between-subjects control. As the participants were not observed in a controlled experimental setting where they recorded their messages all within the same time period, it was crucial to construct a well-defined control group that would account for the confounding factors influencing our outcome – MHE in private messages – as much as possible. Using the set of participants that did not receive any OH messages, we created a first control group that represented those non-OH participants who had interacted on Instagram by sending and/or receiving private messages during a similar time period as the OH participants.

Then for each of these non-OH participants, we created a placebo OH message to construct the pre-OH and post-OH sets of messages (of duration $n = 7, 14, 21, 28$ days each) similar to those for the OH participants. Although constructing a control group of pre-OH and post-OH messages around a placebo message that was sent on the exact same day as those from the OH participants would have been ideal, it was not feasible due to resultant sparsity. Therefore, we controlled not on the date level but on the season level; this achieved the purpose of controlling for temporal factors such as seasonal effects on mental health, while maintaining a reasonable sized dataset for analysis. We assigned seasons to the timestamp of each OH message and constructed placebo messages in the non-OH set that matched the season of the corresponding OH messages of the treatment group.

In addition, we utilized a pre-trained BERTopic model (Grootendorst 2022) to label each conversation with a topic label–this ensured that we constructed a control group of messages not only within the same seasonal period but also matching topic of the conversations. For this, for each given OH message, we first gathered all the messages of each non-OH user that fell within the same season as the OH message. Then we constructed matched pre-OH and post-OH pairs of control messages by selecting those conversations of the non-OH participants whose BERTopic labels were the same as those of the pre-OH and post-OH messages of each treatment user. Since many conversations were likely to be not topically similar, we used a sliding time window of $n = 7, 14, 21, 28$ days within the list of messages of the non-OH participants to get adequate matched pairs for each treatment instance. We denote this group as the *between-subjects control* group from hereon.

To assess the quality of matching obtained through the above BERTopic model, we used the effect size (Cohen's d) metric to quantify the standardized differences between the treatment and the between-subjects control group. We computed the effect size for each time window by measuring the mean difference in the word frequency of the topics between the treatment and the between-subjects control group. The Cohen's d for each of the time windows ranged 0.0017 to 0.0016, demonstrating that there was minimal difference between the two groups, indicative of better matching.

Within-subjects control. One's lived experiences affect one's perception of what consists online harassment and the impact it may have on their subsequent mental health. Therefore, we needed a control group comprising the same group of OH users, wherein we would assess changes in our outcome in response to a received placebo OH message, from before they received any online harassment message. For any given pre-OH/post-OH pair of messages of an OH participant, we gathered a new set of pre-OH/post-OH messages that were from a time period before the treatment (received OH message) and had no overlap with the $n = 7, 14, 21, 28$ day time periods of any other set of pre-OH/post-OH messages of the said participant. We denote this control group as *within-subjects control*.

Category	Description
Treatment	Messages sent by the online harassment victims that are within a given time frame pre/post an online harassment message
Within-subjects control	Messages that sent by online harassment victims that are not close to any online harassment messages
Between-subjects control	Messages sent by people that did not receive online harassment, sent in the same season of those in the treatment group

Table 3: Brief description of each group of messages used to construct the treatment and control groups.

Category	Time Window	
	7	14
Pre-OH	1,268	1,820
Post-OH	1,299	1,973
Within-Subj. Control Pre-OH	15,722	15,982
Within-Subj. Control Post-OH	15,278	15,461
Between-Subj. Control Pre-OH	314	313
Between-Subj. Control Post-OH	289	265

Table 4: Message count corresponding to the $n = 7, 14$ day durations preceding and succeeding a real or placebo OH in the treatment and control groups.

Table 3 summarizes each group of messages established above. Table 4 gives descriptive statistics for each group.

Difference-in-Differences (DID) Analysis Approach

Difference-in-Differences (DID) analysis is a traditional quasi-experimental research approach which aims to examine the difference between the changes in the outcomes pre and post treatment versus those of a control group (Angrist and Pischke 2009). DID analysis aims to compare the convergence pattern of the treatment group to that of the control group, allowing us to identify the mental health impact of online harassment. This enables us to understand the association between online harassment while controlling for underlying trends in the mental health expression in messages that are unrelated to the existence or absence of having received online harassment (Dimick and Ryan 2014).

Once we obtained the raw MHE scores in each message using the mental health classifiers described in the previous

section, we converted each score (in the [0,7] range) to a z-score based on the mean and standard deviation of all the messages in the associated pre-OH or post-OH period, corresponding to the real or placebo OH message.

We averaged the MHE z-scores of each pre-OH and post-OH period to obtain a single z-score for each pre-OH and post-OH period. As a result, corresponding to a single (real or placebo) OH message, we obtained six MHE z-scores: pre-OH/post-OH MHE from the treatment group, pre-OH/post-OH MHE from the between-subjects control, and pre-OH/post-OH MHE from the within-subjects control. We then computed the following to assess the change in MHE within each of the three groups:

$$D_p = z_a(p) - z_b(p) \tag{1}$$

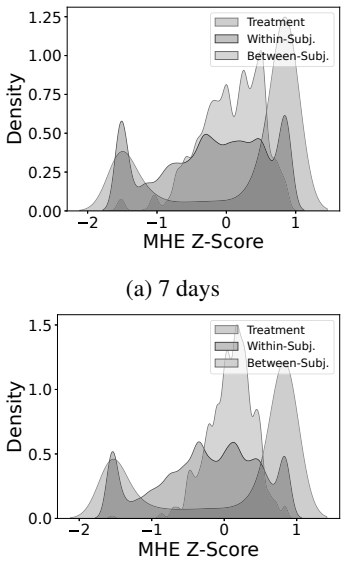
where $z_b(p)$ and $z_a(p)$ respectively represent the averaged MHE z-scores corresponding the pre-OH and post-OH messages of group p, where groups could be treatment (tr), between-subjects control (bc), or within-subjects control (wc). Using the three differences D_{tr}, D_{bc}, and D_{wc}, we then measured the DID of the treatment group with respect to each of the control groups–more specifically, $DID_{bc} = D_{tr} - D_{bc}$ and $DID_{wc} = D_{tr} - D_{wc}$. In simple terms, these DIDs quantify if the change in MHE z-score during the period following receipt of an OH message in the treatment group is greater than the change observed in either of the controls. This was performed for each of the four time windows: $n = 7, 14, 21, 28$ days.

Findings of the DID Analysis

Days	D_{tr}	D_{wc}	D_{bc}	$D_{tr} - D_{wc}$	$D_{tr} - D_{bc}$
7	0.1542	-0.31312	-0.00096	**0.46732**	**0.15516**
14	0.11036	-0.28799	0.02534	**0.39835**	**0.08502**
21	-0.00958	-0.27335	0.01196	**0.26377**	-0.02154
28	0.00910	-0.18009	0.01802	**0.18919**	-0.00892

Table 5: Results of the DID analysis.

Through the DID analysis, for the treatment group, we found that the difference between the post-OH and pre-OH MHE scores (D_{tr}) was uniformly positive than those of both of the control groups (D_{wc}, D_{bc}) across $n = 7, 14$ day time windows. Further, from Table 5, for the 7 day window, we found that the DID between the treatment and both the controls was positive, indicating that the increase in MHE for the treatment group post-OH was more than any changes in MHE observed for either of the controls. Specifically, the DID for the within-subjects control (DID_{wc}) was relatively higher ($0.154 - (-0.313) = 0.467$ for $n = 7$ days and $0.110 - (-0.287) = 0.398$ for $n = 14$ days) than that between the treatment and the between-subjects control (DID_{bc}): $0.154 - (-0.00096) = 0.155$ for $n = 7$ days and $0.110 - 0.025 = 0.085$ for $n = 14$). The positive DID between the treatment group and the control groups persists for the time window of 14 days as well. For time windows of 21 and 28 days, we observe that although the DID value between the treatment group and the within-subjects control group remained positive and demonstrates a gradually declining pattern, the positive DID between the treatment group and the between-subjects control seems to diminish past 14 days. In subsequent analysis, we focus on the first two time windows, $n = 7, 14$ where the mental health impact of harassment seems most pronounced.



(a) 7 days

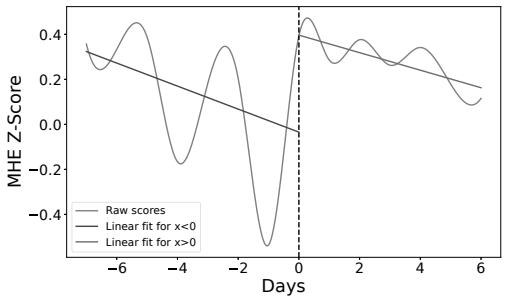
(b) 14 days

Figure 2: KDE of the MHE z-scores between the treatment, within-subj., and between-subj. control groups.

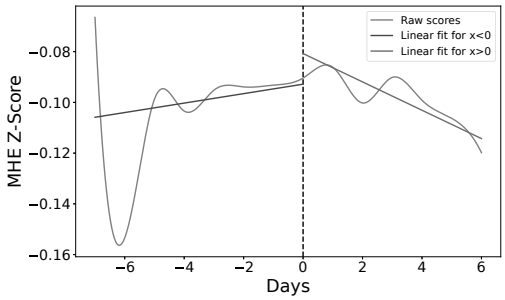
These DID findings are further bolstered by the kernel density distribution graph of the MHE z-scores of each group. Shown in Figure 2, we observe that the treatment group's MHE scores are more skewed to higher values (mean: 0.21, 0.1), relative to both controls mean: -0.21, -0.24 for within and 0.06, 0.09 for between-subjects control), which show a relatively even distribution. This aligns with our findings above, where the magnitude of change in both of the control groups were smaller than that of the treatment, indicating the harassment incident to have a greater impact (by 11%-200%) on the mental health of the victims than in cases where was no harassment.

Unpacking these results further, we generated a temporal graphs of the MHE z-scores of the treatment, within-subjects control, and between-subjects control groups to examine relative differences in their 7- and 14 day-trends surrounding the receipt of a real or placebo OH message: see Figures 3 and 4. We used spline interpolation to illustrate the change of average z-scores and also fit a linear model to show the trend of the z-scores. The treatment group showed an increase in both the slope and intercept (-0.03897-(-0.05114)=0.01217 for slope and 0.36941-(-0.0346)=0.43102 for intercept) for the time window of $n = 7$. The within-subjects control and between-subjects control both showed a decrease in the slopes (-0.00558-0.00186=-0.00744 for within-subjects control and -0.00133-0.00255=-0.00389 for between-subjects control) for the 7-day window. Here, the intercept slightly increased for the within-subjects control (-0.08076-(-0.09283)=0.01207) while the between-control showed a decrease (0.06997-0.07928=-0.0093). For the case where $n = 14$, the slope of the treatment group exhibited a negative change (decrease from -0.0313 to -0.00901), while the intercept increased (0.30791 to 0.35694). These patterns were not present in the two controls for $n = 14$.

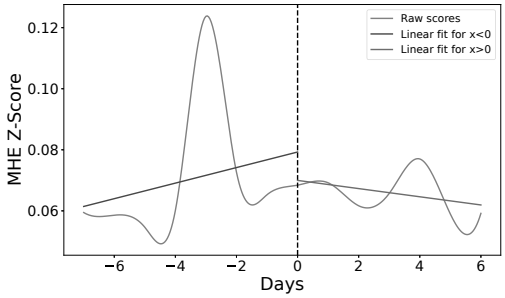
Overall, we note that the absolute magnitude of change for both $n = 7, 14$ was greater in the treatment group compared to both control groups. In Figure 3, a sharp increase in MHE z-scores was ob-



(a) Treatment; 7 days

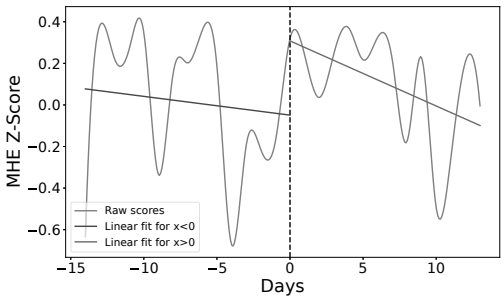


(b) Within-subj; 7 days

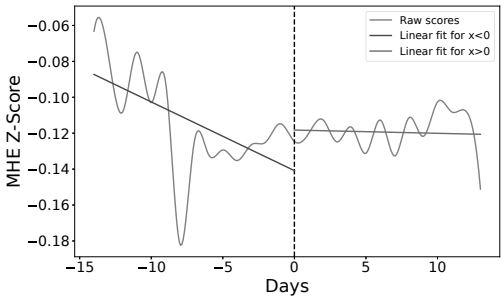


(c) Between-subj; 7 days

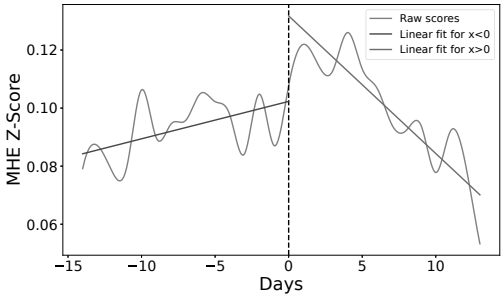
Figure 3: Temporal graph of the MHE z-scores of messages of the treatment, within-subjects control, and between-subjects control groups for a time window of 7 days.



(a) Treatment; 14 days



(b) Within-subj; 14 days



(c) Between-subj; 14 days

Figure 4: Temporal graph of the MHE z-scores of messages of the treatment, within-subjects control, and between-subjects control groups for a time window of 7 days.

served in the treatment group on the day before throughout the day after the online harassment message, whereas both control groups showed a comparatively smaller fluctuation of MHE z-scores. We also note the general trend of decrease in the treatment group following the initial sharp increase for both $n = 7, 14$.

Assessing Validity of the DID Analysis

To examine whether the above observed DIDs between the treatment and the two control groups is statistically significant, drawing on prior research (Bertrand, Duflo, and Mullainathan 2004), we trained an Ordinary Least-Squares (OLS) Regression Model on the MHE scores of the pre-OH and post-OH messages associated with each real or placebo OH message in the three groups. Each MHE score was associated with three variables: g (whether it corresponded to the treatment or one of the control groups), t (whether it corresponded to the pre-OH or post-OH period), and categorical variables based on the OH message ID that were used as dummy variables. This OLS approach would enable us to establish that the observed difference (DID) between the MHE changes in the treatment group and those in the two control groups did not occur out of random chance but rather due to the treatment, which in this case are the receipt of the OH messages. The regression model was fitted as follows:

$$Y = \beta_0 + \beta_1 * g + \beta_2 * t + \beta_3 * (g * t) + \Sigma_{i=1}^{n} r_i \qquad (2)$$

where the variable r represents the categorical dummy variables. The β coefficients each were associated with the dependent variables used in fitting the regression model. Figure 5 shows the visual representation of the coefficients in relation to the MHE of the treatment and control groups–when we visualize the MHE scores for the treatment and the control groups, the green line represents the MHE scores of the control group. The vertical line depicts the time of the treatment–in this case, an online harassment message–that would affect the MHE scores. The treatment group's MHE scores, which is shown as the blue line, then would be influenced by the online harassment message and divert from the expected projection of the MHE scores had the online harassment message did not happen; which is shown as the dotted blue line. To that end, we

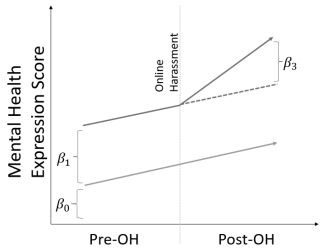


Figure 5: DID graph with regression model coefficients.

focused on β_3, a coefficient that is dependent on both the time (pre- or post-OH) and the presence (OH present or absent) variables.

Applying this OLS model to our data, we found that the impact of OH on the MHE outcomes were statistically significant (F-stat: $1.78^{33}, 6.83^{32}; (p < .05)$ and effect size of $\beta_3 : \eta^2 = 0.85, 0.51; (p < .05)$ for $n = 7, 14$) giving validity to the DID observations that OH incidents were followed by heightened MHE scores in a period 7-14 days.

Comparison with Counterfactuals

To further establish the causal relationship between the observed changes in the MHE z-scores to the online harassment messages, we adopted a multi-prong counterfactual comparison approach from ElSherief et al. (2021)'s study.

A Synthetic Counterfactual We implemented a permutation test to rule out the possibility that the observed temporal changes in the MHE z-score happened by chance (Anagnostopoulos, Kumar, and Mahdian 2008). Essentially, for each set of MHE z-scores of a online harassment message, we conducted 1000 permutations to obtain a set of synthetic counterfactuals; then measured the average difference of the MHE z-scores (D_{tr}) over a 7-day and a 14-day pre-/post-OH period. For these synthetic counterfactuals, we found that the change in the permuted MHE z-scores in the post-OH period with respect to before, was significantly smaller (0.16187 and 0.11127 for 7 and 14-days respectively; $p < 10^{-15}$) than that in the actual treatment data, by 104.97% and 100.82%. The mean, variance, and the standard deviation for the 7 and 14 day time windows are shown in Table 6.

Window	Avg. Change in MHE	Variance	Std. Dev.
7 days	-0.00767	0.00263	0.05127
14 days	-0.00091	0.00101	0.03185

Table 6: Outcomes of the permutation test showing the mean change in MHE z-scores in the synthetic counterfactuals.

A Simulated Counterfactual Using the MHE z-scores of the pre-OH period, we inferred the post-OH MHE z-scores assuming that the treatment–in our case, the online harassment message–did not occur. This allowed us to understand what the MHE z-score would have looked like had the online harassment message not been received by a particular participant. Our rationale is that if there was significant difference between the inferred values and the actual values, we could conclude that the change in MHE z-scores was due to the online harassment message. To infer the post-OH MHE z-scores, we trained an auto-regressive integrated moving average (ARIMA) model (Box et al. 2015). Using the ARIMA model, we then forecasted the values of the post-OH and measured the Root Mean Square Error (RMSE) between the forecasted values and the actual MHE z-scores from the post-OH period. We observed the average RMSE between the forecasted and actual values to be 0.87 and 1.06 for time windows of 7 and 14, respectively. The forecasted values were 0.18 and 0.33 (79.31% and 221.21%) lower than that of the actual MHE z-scores in the treatment data. A two-sample t-test ($t = 2.8; p < .005$ and $t = 5.18; p < 10^{-7}$ for 7 and 14 days respectively) further supported our observation that the inferred values were significantly different from the actual MHE z-scores.

7-day time window		
Positive LLR		
Unigram	Bigram	Trigram
clueless,5.95	embarrassed because,5.96	hurt do not,5.99
embarassed,5.90	blocked him,5.94	can not anything,5.97
cryyyyy,5.90	really mad,5.91	feeeell get pissed,5.92
Negative LLR		
omggg,-5.13	talking bout,-4.78	did not hear,-4.80
learn,-4.03	time video,-4.73	like valid do,-4.67
wisdom,-3.43	know say,-4.17	sorry like want,-4.01
14-day time window		
Positive LLR		
clueless,5.98	life support,7.92	can not anything,6.07
idgaf,5.93	hate this,6.07	i am upset,6.04
idccc,5.88	fucking die,5.99	want to kill,6.02
Negative LLR		
Unigram	Bigram	Trigram
lmfaoo,-6.31	stop i,-5.72	am scared because,-4.67
wassup,-4.17	not talking,-4.86	am sorry love,-4.65
walking,-3.93	talking shit,-4.70	what is wrong,-3.72

Table 7: n-grams with positive and negative LLR.

Posthoc Contextualization of the Findings

To examine our causal findings in context, we extracted the $n = 1, 2, 3$-grams from the messages sent in the $n = 7, 14$ days after the online harassment message and calculated their Log Likelihood Ratio (LLR) with respect to their occurrences in the messages sent before the online harassment message. A higher LLR indicates that the n-gram is more frequent in the post-OH period while a lower value shows greater frequency pre-OH. We then examined $n = 1, 2, 3$-grams that had at least an absolute magnitude of 3 or higher.

Per Table 7, post-OH messages contained terms that exhibit helplessness, hopelessness, and distress such as *cryyyyy*, *clueless*, and *i am upset*; these phrases indicate lower mood and declined cognitive processing, both of which have been observed to be associated with elevated feelings of mental health concerns (De Choudhury, Counts, and Horvitz 2013). Note that there were also some terms with negative LLR that contained negative sentiment (e.g., *am scared because*, *am sorry love*), but in a different manner; for example indicating fear or apology.

Discussion

New Benchmarks for Studying Private Interactions

Theoretically-grounded measurement and validation of MHE in private networked spaces. Considering how self-disclosure and seeking for help in private conversations provide support and aid in the overcoming of traumatic events (Andalibi et al. 2018), it is important to quantify how people express their mental statuses in these private networked spaces. Yet, as mentioned earlier, there has been limited understanding about how

people express their mental state in private messages. Our work, inspired by the approach from previous literature (Saha et al. 2019), provides a valid benchmark for measuring mental health in private messages using (linguistically equivalent) public data on self-disclosures of mental illnesses. A notable strength of our work is that we validated our approach using participants' assessments of their mental health, recorded via standardized questionnaires. This validation approach is more theoretically grounded compared to prior approaches where self-reported diagnoses (e.g., "I am suffering from depression") were taken as standalone ground truth or where third parties annotated social media postings for mental health expression. Ernala et al. (2019) noted these approaches to suffer from construct validity issues, and advocated for theoretically-grounded mental health measurement. Here, we achieved this, in the context of users of private networked spaces through a range of psychometric instruments. Since we further showed distinct mental health expressions of harassed and non-harassed individuals (with the former experiencing worsened mental health), we provide additional face validity to our approach of assessing mental health concerns in private networked spaces.

Advancing causal investigations of impact of harassment in private channels. Our approach offers a new way to measure the mental health impact of online harassment, addressing a gap in the current research. Rather than just focusing on detecting incidents of harassment, as has been the focus of prior work, we analyzed how victims are affected by these events. Our study found that victims exhibit worsened mental health in private conversations for at least a short period of time following a harassment incident, and that this effect spills over to other conversations beyond the one involving the perpetrator. This demonstrates the direct impact of harassment on one's language use and style. In this way, we provide a benchmark for assessing the mental health impact of online harassment and can inform the development of better detection systems and support for victims.

Our work extends prior research by using a causal inference approach to measure impact. Past research has largely relied on retrospective self-reports or correlations, which may be subject to recall bias or confounding effects and may not account for baseline mental health status (De Choudhury 2013). Cross-sectional designs are often limited in being able to establish causality and quantify change. Our DID and counterfactual modeling approach addressed these issues by utilizing a large, longitudinal dataset of Instagram conversations with self-reported instances of harassment.

Broadly, we were not only able to establish that online harassment can lead to worsened mental health, we also examined this impact by looking at varied durations, ranging from one to four weeks. As perhaps expected, the causal effect was the most pronounced in the immediate week, and showed a monotonic decline thereafter. We know from clinical observations that individuals are resilient to the influence of traumatic events and can recover from such event as time passes (Butler et al. 2009). Our findings are consistent with this finding, since we see that past two weeks following experience of harassment, the heightened mental health impact tends to wash away. While this shows that the impact may be short-lived, but we note that the impact during this short period can be notably distinct from the same individual's baseline mental health, or compared to similar individuals who were not harassed online. Our study thus sets a new benchmark in quantifying the impact of harassment, and future research may further investigate this across platforms, forms of harassments, diverse populations, and so on.

Finally, our findings confirm connections between online harassment and emotional, psychosomatic problems, social difficulties, and psychological safety (Duggan 2014; Brody 2021; Aponte and Richards 2013). The impact of online harassment on mental health may lead to a vicious cycle where harassment worsens mental health and poor mental health further victimizes individuals in future incidents (Arseneault, Bowes, and Shakoor 2010). This vicious cycle can be particularly threatening in private networked spaces that are often un- or under-moderated, and where the largely dyadic form of interactions may mean bystanders are unavailable to confront perpetrators or provide help to victims in need. Future research could leverage our approach to measure mental health impacts as a benchmark and investigate the nature of this vicious cycle to inform better design of private networked spaces as well as possible interventions.

Design Implications for Private Networked Spaces

A restorative justice approach to tackling mental health harms of harassment. There are almost two decades of research designing automated machine learning tools for harassment detection online (Kim et al. 2021a), and scholars have repeatedly emphasized the need for interventions (e.g., moderation of perpetrators, deletion of harassing messages, community/bystander reporting mechanisms) inspired by such automated systems to curb harassment from happening repeatedly from the same or similar perpetrators. While these efforts are commendable, all machine learning models are inherently uncertain, and thus detection algorithms, even today, are far from being perfectly accurate in all contexts, populations, or platforms. With such systems failing to deliver in certain occasions, victims might be "left out in the cold" to cope alone without support, perhaps even with the perpetrator still present in the networked space without facing any real consequences. Even when the offenders are "punished" by platforms (e.g., in the form of banning or content-removal), such a retributive approach often leaves out the victim from the delivery of justice, offering them little to no agency in choosing what punishment might be the best for their wellbeing. Such neglect may be perceived to be not just unjust, but our work reveals that it could have significant negative repercussions due to victims expressing aggravated mental health concerns following such an often unpleasant, sometimes harmful experience. Thus, by quantifying the mental health impacts of harassment, we may be able to mediate and provide support to the victims. These interventions can take several forms, and ethical questions abound as to what appropriate victim-centered ways to engage in private networked spaces could be.

Our suggestions for design here draw upon a restorative justice framework (Van Ness and Strong 2014), where the idea is "to get offenders to take responsibility for their actions, to understand the harm they have caused, to give them an opportunity to redeem themselves, and to discourage them from causing further harm." Social media platforms that include private channels of communication may consider strategies where perpetrators of harassment are given an opportunity to justify their actions and adopt steps that could repair their often dyadic private interaction with the respective victims. Victims, complementarily, could be provided resources for coping and social support, on the platform (e.g., giving an option to connect with a close social tie with whom active private interactions are present) or elsewhere (e.g., virtual therapy). Safe spaces with self-reflection or conversational features may also be created for the victims to share their experiences. In addition, knowing that mental health impacts of harassment exist, victims could nominate bystanders in their private channels who, with appropriate consent, may serve to mediate future instances of harassment with an eye to minimizing mental health harms.

Sensitivity to mental health impacts: a transformative justice approach to tackling online harassment. Along similar lines as above, moderation and interventional efforts to address online harassment could consider the impact on victims'

mental health central to platform policies. This could be akin to a transformative justice framework (Nocella and Anthony 2011), that allows investigating the root causes of injustice and uses a harms reduction approach that seeks to "lessen the negative social and/or physical consequences associated with various human behaviors." Most harassment detection techniques do not consider the impact on the victim as an indicator of the presence of severity of the incident. We suggest that with a transformative justice lens, moderation and intervention strategies could build upon our findings to not only detect the contextual features of a given point in time when/where harassment occurs, but also consider an assemblage of features – e.g., specific language present in messages or a sudden negative change in social media behavior – relating to the possible harms among victims. This would allow for a victim-centered approach to identifying what constitutes harassment, when, and how. It would also allow for better unpacking of the consequences and implications of different moderation techniques, ensuring that evaluation of these techniques considers the mitigation of victims' mental health harms as key to success. A transformative justice lens may also enable new interventions that involve education and awareness campaigns promoting digital literacy, healthy relationships, and consent culture in private social media conversations.

Conclusions, Limitations, and Future Work

We explored how online harassment incidents influence the mental health of victims expressed in private conversations. We developed a robust causal inference framework that utilized rich data spanning 1.4 million private messages shared by 80 youth on Instagram. In comparison with two carefully constructed control groups, a difference-in-differences and a counterfactual modeling analysis revealed the treatment group (those who experienced self-identified harassment in Instagram conversations) to exhibit increased mental health concerns. These concerns persisted for a short period of time (7 and 14 days) following the harassment incident and spilled over to multiple conversations, manifesting in varied linguistic characterizations indicative of aggravated mental health. Future work can explore the extent to which these effects might still have a more prolonged footprint, whether the effects could have individual differences given the lived experience of the victim, whether the identity of the perpetrator of harassment has any role in lessening or exacerbating the observed impact, and the extent to which our findings may extend to social media platforms beyond Instagram. Although the focus on a single platform and the use of transfer learning to assess mental health expressions could be perceived as limitations of this work, through this formative study we hope to inspire others to recognize the effect of online harassment on the expressed mental statuses of victims in private networked spaces, considering which has the potential to improve future online harassment detection systems as well as to devise ethical and just ways victims of harassment could be better supported.

Acknowledgements

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

References

Adams, M.; Scarduzio, J. A.; Limperos, A. M.; and Fletcher, J. 2022. Younger employees' sexual harassment experiences on Facebook's public feed versus direct messages: How the online setting impacts uncertainty and coping. *Sexuality & Culture*, 26(5): 1840–1857.

Almerekhi, H.; Jansen, B. J.; and Kwak, H. 2020. Investigating toxicity across multiple Reddit communities, users, and moderators. In *Proc. WebConf*, 294–298.

Anagnostopoulos, A.; Kumar, R.; and Mahdian, M. 2008. Influence and correlation in social networks. In *Proc. KDD*, 7–15.

Andalibi, N.; Haimson, O. L.; Choudhury, M. D.; and Forte, A. 2018. Social support, reciprocity, and anonymity in responses to sexual abuse disclosures on social media. *Proc. TOCHI*, 25(5): 1–35.

Angrist, J. D.; and Pischke, J.-S. 2009. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

Aponte, D. F. G.; and Richards, D. 2013. Managing cyberbullying in online educational virtual worlds. In *Proc. IE*, 1–9.

Arseneault, L.; Bowes, L.; and Shakoor, S. 2010. Bullying victimization in youths and mental health problems:'Much ado about nothing'? *Psych med*, 40(5): 717–729.

Bertrand, M.; Duflo, E.; and Mullainathan, S. 2004. How much should we trust differences-in-differences estimates? *The Quarterly journal of economics*, 119(1): 249–275.

Box, G. E.; Jenkins, G. M.; Reinsel, G. C.; and Ljung, G. M. 2015. *Time series analysis: forecasting and control*. John Wiley & Sons.

Brody, N. 2021. Bystander Intervention in Cyberbullying and Online Harassment: The Role of Expectancy Violations. *International Journal of Communication*, 15: 21.

Butler, L. D.; Koopman, C.; Azarow, J.; Blasey, C. M.; Magdalene, J. C.; DiMiceli, S.; Seagraves, D. A.; Hastings, T. A.; Chen, X.-H.; Garlan, R. W.; et al. 2009. Psychosocial predictors of resilience after the September 11, 2001 terrorist attacks. *J. Nervous and Mental Disease*, 197(4): 266–273.

Cañas, E.; Estévez, E.; Martínez-Monteagudo, M. C.; and Delgado, B. 2020. Emotional adjustment in victims and perpetrators of cyberbullying and traditional bullying. *Social Psychology of Education*, 23(4): 917–942.

Chatzakou, D.; Kourtellis, N.; Blackburn, J.; De Cristofaro, E.; Stringhini, G.; and Vakali, A. 2017. Mean birds: Detecting aggression and bullying on twitter. In *Proc. WEBSCI*.

De Choudhury, M. 2013. Role of social media in tackling challenges in mental health. In *Proc. SAM*, 49–52.

De Choudhury, M.; Counts, S.; and Horvitz, E. 2013. Social media as a measurement tool of depression in populations. In *Proc. WEBSCI*, 47–56.

Dimick, J. B.; and Ryan, A. M. 2014. Methods for evaluating changes in health care policy: the difference-in-differences approach. *Jama*, 312(22): 2401–2402.

Dinakar, K.; Reichart, R.; and Lieberman, H. 2011. Modeling the detection of textual cyberbullying. In *AAAI Workshop - Technical Report*, volume WS-11-02, 11–17. ISBN 9781577355182.

Duggan, M. 2014. Part 4: The Aftermath of Online Harassment — Pew Research Center. https://www.pewresearch.org/internet/2014/10/22/part-4-the-aftermath-of-online-harassment/. Accessed: 2021-11-23.

ElSherief, M.; Saha, K.; Gupta, P.; Mishra, S.; Seybolt, J.; Xie, J.; O'Toole, M.; Burd-Sharps, S.; and De Choudhury,

M. 2021. Impacts of school shooter drills on the psychological well-being of American K-12 school communities: a social media study. *Nature HSSC*, 8(1): 1–14.

Ernala, S. K.; Birnbaum, M.; Candan, K.; Rizvi, A.; Sterling, W.; Kane, J.; and De Choudhury, M. 2019. Methodological gaps in predicting mental health states from social media: triangulating diagnostic signals. In *Proc. CHI*, 1–16.

Grootendorst, M. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.

Hafford-Letchfield, T.; Toze, M.; and Westwood, S. 2022. Unheard voices: A qualitative study of LGBT+ older people experiences during the first wave of the COVID-19 pandemic in the UK. *Health & social care in the community*, 30(4): e1233–e1243.

Harasgama, K. S.; and Jayamaha, S. 2023. Online Harassment in Sri Lanka: A Thematic Analysis. *Social Sciences*, 12(3): 176.

Hemphill, S. A.; Kotevski, A.; and Heerde, J. A. 2015. Longitudinal associations between cyber-bullying perpetration and victimization and problem behavior and mental health problems in young Australians. *International journal of public health*, 60(2): 227–237.

Imbens, G. W.; and Rubin, D. B. 2015. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Kim, S.; Razi, A.; Stringhini, G.; Wisniewski, P. J.; and De Choudhury, M. 2021a. A Human-Centered Systematic Literature Review of Cyberbullying Detection Algorithms. *Proc. CSCW*, 5: 1–34.

Kim, S.; Razi, A.; Stringhini, G.; Wisniewski, P. J.; and De Choudhury, M. 2021b. You Don't Know How I Feel: Insider-Outsider Perspective Gaps in Cyberbullying Risk Detection. In *Proc. ICWSM*, 290–302.

Kumar, A.; and Sachdeva, N. 2019. Cyberbullying detection on social multimedia using soft computing techniques: a meta-analysis. *Multimedia Tools and Applications*, 78(17): 23973–24010.

Lewis, S. C.; Zamith, R.; and Coddington, M. 2020. Online harassment and its implications for the journalist–audience relationship. *Digital Journalism*, 8(8): 1047–1067.

Liberatore-Maguire, E.; Devlin, A.; Fisher, S.; Ramsey, F.; Grunwald, H.; Brownstein, K.; and Morrison, M. 2022. The unseen epidemic: trauma and loneliness in urban midlife women. *Women's Midlife Health*, 8(1): 1–10.

Liu, T.; Meyerhoff, J.; Eichstaedt, J. C.; Karr, C. J.; Kaiser, S. M.; Kording, K. P.; Mohr, D. C.; and Ungar, L. H. 2022. The relationship between text message sentiment and self-reported depression. *J. affective disorders*, 302: 7–14.

Matthay, E. C.; Hagan, E.; Gottlieb, L. M.; Tan, M. L.; Vlahov, D.; Adler, N.; and Glymour, M. M. 2021. Powering population health research: Considerations for plausible and actionable effect sizes. *SSM-population health*, 14: 100789.

McHugh, B. C.; Wisniewski, P. J.; Rosson, M. B.; Xu, H.; and Carroll, J. M. 2017. Most teens bounce back: Using diary methods to examine how quickly teens recover from episodic online risk exposure. *Proc. CSCW*, 1: 1–19.

Murphy, L.; Markey, K.; O'Donnell, C.; Moloney, M.; and Doody, O. 2021. The impact of the COVID-19 pandemic and its related restrictions on people with pre-existent mental health conditions: A scoping review. *Archives of Psychiatric Nursing*, 35(4): 375–394.

Naslund, J. A.; Bondre, A.; Torous, J.; and Aschbrenner, K. A. 2020. Social media and mental health: benefits, risks, and opportunities for research and practice. *Journal of technology in behavioral science*, 5: 245–257.

Nocella, A. J.; and Anthony, J. 2011. An overview of the history and theory of transformative justice. *Peace & conflict review*, 6(1): 1–10.

Rahman, M. S. 2020. The advantages and disadvantages of using qualitative and quantitative approaches and methods in language "testing and assessment" research: A literature review.

Razi, A.; AlSoubai, A.; Kim, S.; Naher, N.; Ali, S.; Stringhini, G.; De Choudhury, M.; and Wisniewski, P. J. 2022. Instagram Data Donation: A Case Study on Collecting Ecologically Valid Social Media Data for the Purpose of Adolescent Online Risk Detection. In *Proc. CHI EA*, 1–9.

Rigby, K. 2003. Consequences of bullying in schools. *The Canadian journal of psychiatry*, 48(9): 583–590.

Rosa, H.; Pereira, N.; Ribeiro, R.; Ferreira, P.; Carvalho, J. P.; Oliveira, S.; Coheur, L.; Paulino, P.; Simão, A. V.; and Trancoso, I. 2019. Automatic cyberbullying detection: A systematic review. *Comp. in Human Behavior*, 93: 333–345.

Saha, K.; Chan, L.; De Barbaro, K.; Abowd, G. D.; and De Choudhury, M. 2017. Inferring mood instability on social media by leveraging ecological momentary assessments. *Proc. IMWUT*, 1(3): 1–27.

Saha, K.; and De Choudhury, M. 2017. Modeling stress with social media around incidents of gun violence on college campuses. *Proc. CSCW*, 1: 1–27.

Saha, K.; Sugar, B.; Torous, J.; Abrahao, B.; Kıcıman, E.; and De Choudhury, M. 2019. A social media study on the effects of psychiatric medication use. In *Proc. ICWSM*, volume 13, 440–451.

Ståhl, S.; and Dennhag, I. 2021. Online and offline sexual harassment associations of anxiety and depression in an adolescent sample. *Nordic journal of psychiatry*, 75(5).

Stamatis, C. A.; Meyerhoff, J.; Liu, T.; Sherman, G.; Wang, H.; Liu, T.; Curtis, B.; Ungar, L. H.; and Mohr, D. C. 2022. Prospective associations of text-message-based sentiment with symptoms of depression, generalized anxiety, and social anxiety. *Depression and anxiety*, 39(12): 794–804.

Stevens, F.; Nurse, J.; and Arief, B. 2021. Cyberstalking, cyber-harassment, & adult mental health: A systematic review. *Cyberpsy, Behav, & Soc Networking*, 24(6): 367–376.

Uban, A.-S.; Chulvi, B.; and Rosso, P. 2021. An emotion and cognitive based analysis of mental health disorders from social media data. *Future Gen Computer Sys*, 124: 480–494.

Van Ness, D.; and Strong, K. H. 2014. *Restoring justice: An introduction to restorative justice*. Routledge.

VanderWeele, T.; Jackson, J.; and Li, S. 2016. Causal inference and longitudinal data: a case study of religion and mental health. *SP & PE*, 51: 1457–1466.

Yao, M.; Chelmis, C.; and Zois, D. S. 2018. Cyberbullying detection on instagram with optimal online feature selection. In *Proc. ASONAM*, 401–408. Institute of Electrical and Electronics Engineers Inc. ISBN 9781538660515.

Younas, F.; Naseem, M.; and Mustafa, M. 2020. Patriarchy and social media: Women only facebook groups as safe spaces for support seeking in Pakistan. In *Proc. ICTD*.

AAAI ICWSM Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes, see the Data Collection.
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes.
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes.
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes, see the Data Collection.
   (e) Did you describe the limitations of your work? Yes, see Conclusions, Limitations, and Future Work.
   (f) Did you discuss any potential negative societal impacts of your work? No, but we clearly describe the implications of our work as well as the potential limitations and how future researchers could use and build upon our work in Conclusions, Limitations, and Future Work.
   (g) Did you discuss any potential misuse of your work? No, but we clearly describe the implications of our work as well as the potential limitations and how future researchers could use and build upon our work in Conclusions, Limitations, and Future Work.
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes, see the Data Collection.
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes.
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? NA
   (b) Have you provided justifications for all theoretical results? NA
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA
   (e) Did you address potential biases or limitations in your theoretical framework? NA
   (f) Have you related your theoretical results to the existing literature in social science? NA
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA
   (b) Did you include complete proofs of all theoretical results? NA
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? NA
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? NA
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? NA
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? NA
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? NA
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? NA
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? Yes, see Measuring Mental Health in Private Messages.
   (b) Did you mention the license of the assets? NA
   (c) Did you include any new assets in the supplemental material or as a URL? NA
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? No, but the authors were all under the same United States National Science Foundation under grant IIP-1827700 and the work that described in detail of the data collection was appropriately cited.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? NA
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see **?**)? NA
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see **?**)? NA
6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...
   (a) Did you include the full text of instructions given to participants and screenshots? NA
   (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA
   (d) Did you discuss how data is stored, shared, and deidentified? NA