Proceedings of the Twentieth International AAAI Conference on Web and Social Media (ICWSM 2026)

Adverse Mental Health Events As a Trigger for Online Harassment: Evidence from Online Conversations of Youth Experiencing Psychiatric Hospitalizations

Seunghyun Kim,[1] **Michael L. Birnbaum,** [2,3] **Munmun De Choudhury** [1]

[1] School of Interactive Computing, Georgia Institute of Technology
[2] New York State Psychiatric Institute
[3] Columbia University Medical Center
seunghyun.kim@gatech.edu, Michael.Birnbaum@nyspi.columbia.edu, munmund@gatech.edu

Abstract

Online harassment is a pervasive issue that disproportionately impacts individuals with mental health challenges. Yet, modeling under what circumstances an individual might become susceptible to harassment remains underexplored. This study investigates how psychiatric hospitalization may causally impact the likelihood of experiencing online harassment, using over 360,000 Instagram direct messages shared by individuals with and without mental health diagnoses. Our analysis reveals distinct linguistic patterns in harassment directed at mental health patients, with a significant increase in harassment following hospitalization. A Difference-in-Differences analysis highlights the critical two-week period post-hospitalization as a time of heightened risk to harassment. Additionally, we explore the protective role of social support and resilience through Vector Autoregressive modeling, showing that emotional and informational support significantly reduce the likelihood of harassment. These findings emphasize the need for context-aware interventions and trauma-informed platform designs to protect vulnerable individuals online. By understanding the interplay between mental health events, social support, and online harassment, this research offers new directions for developing strategies to foster safer digital environments.

Content Warning: This paper includes mentions and descriptions of online harassment and mental health challenges.

Introduction

Online harassment is a distressing issue with long-lasting repercussions, particularly for younger demographics, including adolescents, youth, and young adults (Brody 2021). Existing computational research has largely focused on approaches to detect instances of online harassment, such as on social media platforms, either for timely content moderation efforts, or to design interventions that can extend support to those victimized (Kim et al. 2021). Scholars from psychology and social science, on the other hand, have investigated the impact of online harassment, employing methods such as retrospective self-reports and interviews (Campbell 2005; Rigby 2003).

However, this far, online harassment research has largely been agnostic to the unique context of the victims (Ozden and Icellioglu 2014) – after all, not everybody experiences online harassment in the same way – a Pew Research report (Lenhart et al. 2015) notes that "Online harassment tends to occur to different groups in different environments with different personal and emotional repercussions." From what is known, age and gender are most closely associated with the experience of online harassment. Deeper consideration of the victims' life experiences and situations are largely absent from our current understanding of online harassment.

In particular, the relationship between an individual's mental health and their vulnerability to online harassment is a growing area of concern in the digital age. As social media and online interactions become increasingly pervasive, understanding how mental health factors influence a person's susceptibility to cyberbullying and online abuse is crucial. Research has shown that individuals with pre-existing mental health conditions, such as anxiety, depression, and low self-esteem, often face stigma, discrimination, and social isolation (Ybarra and Mitchell 2007; Dredge, Gleeson, and De la Piedad Garcia 2014), making them more susceptible to abuse by bad actors. This vulnerability can, in turn, exacerbate their mental health issues, creating a vicious cycle of distress and harm. As increasing numbers of youth with mental health struggles appropriate online platforms for disclosure and support, it is vital to explore how their mental health experience and journey impact the likelihood of becoming a target of online harassment.

This paper aims how mental health conditions may increase an individual's risk of being harassed online. Seeking to offer insights into the mechanisms that underlie this vulnerability, we specifically ask:

- **RQ1:** *What are the key characteristics that differentiate the online harassment received by people diagnosed with a mental illness compared to those received by others?*
- **RQ2:** *Do adverse mental health events precipitate more online harassment?*
- **RQ3:** *Can online social support and an individual's resilience reduce the likelihood of online harassment in times of relative well-being or following adverse events?*

Towards these RQs, we adopted techniques from natural language processing, large language models (LLMs), and causal inference on a dataset of over 360,000 conversational messages from Instagram, voluntarily shared by 134 mental

health patients and other individuals clinically assessed to be not suffering from a mental health condition (control).

First, our analysis revealed significant differences in the linguistic and lexical characteristics of online harassment received by mental health patients compared to by the control group. Mental health patients were more frequently targeted with personal name-calling and offensive language. Next, a Difference-in-Differences (DID) analysis demonstrated that adverse mental health events in the form of a psychiatric hospitalization can be associated with an increased likelihood of online harassment among mental health patients, especially in the first two weeks following hospitalization. Finally, to investigate whether online support and resilience can offset the impact of online harassment, we trained multiple Vector Autoregressive (VAR) models to predict the proportion of non-harassment messages received by both patients and controls. We found that inclusion of resilience scores inferred from an individual's own Instagram messages as well as emotional and information support assessed from their received messages together improved the prediction performance, highlighting the crucial role of support and resilience in mitigating the experience of harassment.

In summary, this study underscores the complex interplay between social support, resilience, and online harassment among those with mental health struggles. We emphasize the need for multifaceted intervention strategies to protect mental health patients in online environments, contributing to the broader discourse on mental health and digital safety.

Related Work

Mental Health and Online Harassment

The relationship between mental health and online harassment has become increasingly prominent in recent years, as social media usage has soared. Extensive research documents how individuals with mental health conditions are more vulnerable to online abuse, exacerbating issues like anxiety, depression, and social isolation (Dredge, Gleeson, and De la Piedad Garcia 2014; Ozden and Icellioglu 2014). Victims of online harassment often face a compounding effect on their mental health, where digital abuse intensifies pre-existing psychological distress (Gualdo et al. 2015). Studies have demonstrated that mental health patients, especially those dealing with low self-esteem or emotional instability, are frequent targets of abuse (Ybarra and Mitchell 2007). However, there is limited work on how specific mental health events, such as psychiatric hospitalization, influence the likelihood of being harassed online.

Previous research has largely explored demographic factors like age and gender as predictors of online victimization, but fewer studies have examined the effect of acute mental health crises on online interactions (Ozden and Icellioglu 2014; Eroğlu et al. 2015; Ronis and Slaunwhite 2019; Aljasir and Alsebaei 2022). Our study seeks to address this gap by investigating how psychiatric hospitalization triggers increased vulnerability to harassment and how this relationship evolves over time. We apply causal analysis techniques, such as Difference-in-Differences (DiD), to assess how hospitalization alters the likelihood of harassment, providing a new perspective on the temporal dynamics of online abuse.

Moreover, the anonymity afforded by digital platforms can embolden perpetrators, making it easier to target individuals when they are most vulnerable, such as post-hospitalization. Research shows that individuals with mental health challenges are often perceived as weaker, which may contribute to increased victimization (Carmen, Rieker, and Mills 1984). Despite the awareness of these patterns John et al. (2018), few studies have explored how adverse mental health events specifically correlate with harassment risk. Our research adds to this understanding by examining how online harassment manifests around the time of hospitalization, a critical period for mental health patients.

Harassment Detection and Support Systems

In response to the growing prevalence of online harassment, several computational approaches have been developed to detect abusive content. Natural language processing techniques have been widely used to classify harmful messages and flag online abuse (Gashroo and Mehrotra 2024; Madhurima et al. 2024; Arshed et al. 2025; Marshan et al. 2025; Todorovic et al. 2025). For example, Ali et al. (2023) applied machine learning classifiers to detect unsafe in Instagram messages, showing that language-based features such as sentiment and specific word usage can be strong predictors. However, many of these classifiers are trained on generic datasets, often failing to account for the unique linguistic patterns found in harassment targeted at mental health patients. Harassment against vulnerable populations may involve more personalized or insidious forms of abuse, which current detection systems may not fully capture.

A limitation in existing harassment detection models is therefore their lack of contextual awareness, particularly in relation to the personal circumstances of the victim. Prior work by Ozden and Icellioglu (2014) suggests that harassment often intensifies when it is personalized, such as using derogatory language specific to a person's mental health status. Therefore, improving harassment detection systems by incorporating context-aware features is critical. Our work builds on these efforts by identifying distinct linguistic markers in harassment messages directed at mental health patients, which could be incorporated into more sophisticated harassment detection tools.

In addition to improving detection, several studies have emphasized the role of social support in mitigating the effects of harassment. Sharma and De Choudhury (2018) highlighted how support from online communities can help victims cope with trauma. Similarly, Kim et al. (2023) found that a robust social network can act as a buffer, reducing the emotional toll of psychological issues. However, there is limited exploration of how social support interacts with mental health events, such as hospitalization, to protect individuals from harassment. Our research seeks to address this gap. Further, resilience, which refers to an individual's ability to recover from adversity, has also been studied as a protective factor against harassment. McLaughlin et al. (2021) demonstrated how expressions of resilience can deter future harassment, signaling emotional strength and non-reactivity. In this study, we explore how resilience, inferred from Insta-

gram messages, helps reduce the likelihood of harassment following hospitalization.

In summary, while advancements have been made in harassment detection and understanding the role of social support, significant gaps remain regarding the experiences of mental health patients. Our study contributes to this area by investigating how hospitalization increases harassment risk, and how social support and resilience can mitigate this harm.

Data

We utilized a unique dataset of Instagram direct message conversations contributed by individuals as part of a U.S. based federally funded study on discovering and understanding mental health and social media use (Nguyen et al. 2022).

Participant Recruitment Approach. For this study named THRIVE, participants were recruited from June 23, 2016, to December 4, 2020. Two distinct groups were recruited: a) individuals clinically diagnosed with mental illnesses (schizophrenia spectrum disorders and mood disorders) and b) verified healthy controls without a diagnosis of mental illness, all within the ages of 15 to 35 years.

Participants with mental illnesses received their diagnoses based on clinical assessments conducted during their most recent episodes. This diagnostic information was extracted from the medical records of participants following informed consent. Recruitment took place at the Northwell Health Zucker Hillside Hospital and collaborating institutions located in East Lansing, Michigan. Exclusion criteria for this group included individuals with an IQ below 70 (based on clinical assessment), those with autism spectrum disorder, or those with substance-induced psychotic disorder.

In addition, healthy volunteers in the same age range of 15-35 years, were recruited from multiple sources. First, some were selected from an existing database of eligible individuals who had previously undergone screening for other research projects at Zucker Hillside Hospital and had consented to be contacted for further research opportunities. Their healthy status was determined through either the Structured Clinical Interview for the Diagnostic and Statistical Manual of Mental Disorders, conducted within the past 2 years, or the Psychiatric Diagnostic Screening Questionnaire (Zimmerman and Mattia 2001). Individuals were excluded if the screening process revealed clinically significant psychiatric symptoms. Additional volunteers were recruited through a web-based student community research recruitment platform at a southeastern university. Lastly, healthy volunteers were also enlisted from collaborating institutions located in East Lansing, Michigan.

To ensure privacy, informed consent was obtained from adult participants and legal guardians of participants below 18. Assent was obtained from minors. The study's data collection was approved by the institutional review board of the relevant institutions. Appendix A gives further details.

Clinical Data. Following consent, the medical history of each participant was gathered, adhering to Health Insurance Portability and Accountability Act–compliant policies. This encompassed a wide range of information, primary and secondary diagnosis codes, the total number of hospitalizations, as well as admission and discharge dates for each hospitalization event. The hospitalized data was gathered from medical records at the time of the consent for the data collection. The hospitalization records were specifically focused on those related to psychiatric reasons.

Category	Full Sample	Filtered Sample
Age (years). mean (SD)	24.73 (5.64)	22.74 (6.92)
Gender		
Male	127 (47.4%)	65 (48.5%)
Female	141 (52.6%)	69 (51.5%)
Race or ethnicity		
African American or Black	83 (31%)	40 (29.8%)
Asian	43 (16%)	22 (16.4%)
White	112 (41.8%)	56 (41.7%)
Mixed race or other	20 (7.5%)	16 (12.1%)
Hispanic	9 (3.4%)	4 (2.9%)
Pacific Islander	1 (0.4%)	0 (0%)

Table 1: Characteristics of the THRIVE dataset.

Social Media Data. All participants who gave their assent/consent were also requested to download and share their data archives from three plaforms – Facebook, Twitter, and Instagram – a feature provided by the respective platforms compliance with European Union's General Data Protection Regulation. Social media data was acquired from all available platforms for participants who had experienced at least one documented hospitalization event within a 6-month period before their most recent hospitalization. This step ensured that the data collected represented the participants' mental health before any symptomatic exacerbation leading to hospitalization.

Dataset Characteristics. In this paper, we utilized participant-contributed social media data from 268 participants (mean age 24.73 ($\pm$5.64) years; male: 47.4%; 65% were mental health patients and 35% were controls). Many of these participants donated data spanning more than one platform – e.g., 14.2% (38/268) of participants had valid data for all 3 platforms. Speaking of individual platforms, 254 (94.8%) had valid Facebook data, 51 (19%) had valid Twitter data, and 134 (50%) had valid Instagram data (360,375 conversational messages); here valid means the downloaded archive was not corrupt, did not have missing files, and had parsable JSON or HTML data.

Relative to Facebook and Twitter, Instagram is a more widely used social media platform among teens (Anderson and Jiang 2018). Hence, the Instagram data from the 134 participants was used in the remainder of the paper. Table 1 shows the demographic characteristics of the dataset. Figure A1 gives the distribution of messages over the patient and control participants. We had 287,389 private messages from the 87 mental health patient participants, and 74,688 from the 47 controls. This data spanned 1-6 years, with an average duration of about 4 years. Appendix A further discusses the scope and representativeness of this data.

Online Harassment and the Experience of Mental Illness (RQ1)

Detecting Online Harassment Messages

To understand the relationship between online harassment and the experience of a mental illness, we devised an approach to identify online harassment in the social media trace data of the participants. There were no participant-based ground truth labels on online harassment that were available to us, and prior research has found that third party annotations of harassment and bullying may not sufficiently and accurately capture the lived experience of the victims (Kim et al. 2021). Consequently, we employed an online harassment classifier trained on a different dataset of Instagram direct messages of youth (Razi et al. 2022). Razi et al. (2022) gathered this data using an approach similar to the above described THRIVE dataset; their participants' age ranged between 13-18 which significantly overlaps with that of the THRIVE participants. Importantly, in Razi et al. (2022)'s data, the participants themselves labeled their direct messages for their lived experiences of harassment.

The classifier was developed in the following manner. Following Kim et al. (2024), it utilized a list of linguistic features including Linguistic Inquiry and Word Count (LIWC), n-grams, sentiment (based on VADER), BERT, a hate lexicon by Saha, Chandrasekharan, and De Choudhury (2019), and InferSent (Conneau et al. 2017) to provide a binary classification label for each Instagram message. The participant's own annotations of the messages served as ground truth in this annotation task, thus capturing the naturalistic perspectives of the victims directly.

A number of different supervised machine learning models were tested with these features, including Linear Support Vector Machine, Random Forest, Logistic Regression, and Long-Term Short-Term Memory (LSTM). For evaluating the performance of each model, we use the F1-measure, the area under the receiver operating characteristic curve (AUC), and class-specific precision and recall. k-fold cross validation and testing on a 20% heldout dataset revealed the best performing classifier to be LSTM (AUC=0.86; F1=0.87). Specifically, for this LSTM model, we implemented a single-layer architecture with 128 hidden units, followed by a dropout layer (rate = 0.5) to mitigate overfitting, and a fully connected softmax output layer for binary classification. All linguistic features – including LIWC categories, *n*-grams, sentiment scores, hate lexicon features, InferSent embeddings, and contextual embeddings from BERT – were concatenated into a unified feature vector that served as input to the LSTM. Hyperparameter optimization was conducted using grid search with 5-fold cross validation on the training data, exploring learning rates (1e-4 to 1e-2), batch sizes (32, 64, 128), dropout rates (0.3–0.6), and optimizer variants (Adam, SGD with momentum).

The LSTM model was applied to machine label all of the Instagram messages in the THRIVE dataset.

Validity of the Online Harassment Classifier

To assess cross-dataset validity, a random sample of 100 messages were selected; one author then manually labeled each of the messages in a binary fashion to compare with the predictions from the online harassment classifier. The author's labels were corroborated with another coauthor, which yielded a high agreement between classifier and the manual annotations (Cohen's $\kappa = 0.7$). The LSTM based classifier identified a total of 58,230 online harassment messages and 229,159 non-online harassment messages from the mental health patient group, and 13,671 and 61,017 respectively, from the control group's Instagram messages.

LIWC Category	Example Terms	t	p
social	talk, friend, meet	13.923	5.17E-44
article	the, a, an	9.791	1.26E-22
second person	you, your, yours	9.634	5.91E-22
bio	life, live, body	8.828	1.09E-18
swear	damn, hell, shit	7.507	6.07E-14
sexual	sex, love, kiss	6.701	2.07E-11
anger	hate, kill, piss	6.696	2.15E-11
body	hands, feet, blood	6.436	1.23E-10
preposition	to, with, in	6.119	9.41E-10
relative	near, far, above	5.475	4.37E-08

(a)

LIWC Category	Example Terms	t	p
money	cash, money, rich	2.723	0.00645
first person singular	I, me, my	2.497	0.01251
discrepancies	should, would, could	-2.170	0.03001
death	bury, kill, die	-2.222	0.02625
see	view, saw, look	-2.994	0.0027
auxiliary verbs	am, will, have	-3.175	0.00149
tentativeness	maybe, perhaps, guess	-4.570	4.8E-06
future tense	will, going, shall	-4.714	2.4E-06
insight	think, know, consider	-4.749	2.1E-06
third person	they, them, their	-5.479	4.2E-08

(b)

Table 2: Top (a) and bottom (b) 10 LIWC categories for online harassment messages received by mental health patients compared to those received by the health control group.

Linguistic Comparison

Psycholinguistic Differences. After identifying each message containing online harassment, we performed a detailed linguistic analysis using LIWC to assess the linguistic differences between the online harassment messages received by mental health patients and the control group. From each message flagged as containing online harassment, we extracted a 50 dimensional vector, based on the categories of LIWC, including the use of negative emotions (e.g., anger, anxiety), cognitive processes (e.g., insight, causation), and social words (e.g., family, friends). The results were then compared between the two groups, with key metrics of interest including the frequency of negative emotion words and the complexity of cognitive processes in the harassment messages. We implemented an unpaired t-test with False Discovery Rate (FDR) correction for each of the LIWC categories to compare the the two groups so that we analyzed the differences between two independent groups while controlling for multiple comparisons.

Word	SAGE
dante	3.941
kellen	2.946
huncho	2.691
audio	2.456
yoga	2.405
lmfao	2.315
nigga	2.185
tori	2.164
niggas	2.148
kno	2.113

(a)

Word	SAGE
hugo	5.01
college	1.901
ryan	1.672
dating	1.611
classes	1.585
hahahaha	1.398
holy	1.320
kind	1.224
likes	1.157
says	1.157

(b)

Table 3: Distinctive words in the online harassment messages received by the (a) patient and (b) control groups.

Table 2a gives the top 10 LIWC categories where mental health patients received notably higher levels of online harassment. The "social" category exhibited the highest t-statistic (13.923) with a low p-value (10^{-44}), indicating that harassment messages directed at mental health patients tend to adopt strategies of interpersonal interactions, relationships, emotions, and social behavior. Other prominent categories include "article" (t = 9.791, p = 10^{-22}), "second person" (t = 9.634, p = 10^{-22}), and "bio" (t = 8.828, p = 10^{-18}), suggesting that information is conveyed with clarity and specificity, in a conversational style with others, and often focusing on health related topics. Negative emotional content was also prevalent, as indicated by a high t in categories like "swear" (t = 7.507, p = 10^{-14}), "sexual" (t = 6.701, p = 10^{-11}), and "anger" (t = 6.696, p = 10^{-11}).

Conversely, Table 2b presents the bottom 10 LIWC categories where online harassment messages received by mental health patients were significantly less frequent compared to those by the control group. Categories such as "third person" (t = -5.479, p = 10^{-08}), "insight" (t = -4.749, p = 10^{-06}), and "future tense" (t = -4.714, p = 10^{-06}) were lower in messages to mental health patients, indicating a reduced attention to other individuals, a lack of insight and future orientation in thinking. Additionally, the "tentativeness" category (t = -4.570, p = 10^{-06}) and "auxiliary verbs" (t = -3.175, p = 0.00149) were significantly less prevalent, suggesting a style that involves expressing ideas or statements with greater certainty and confidence, avoiding language that suggests doubt, ambiguity, or hesitation in the harassment messages aimed at mental health patients. Other categories with significant differences include "see" (t = -2.994, p = 0.00275) and "discrepancies" (t = -2.170, p = 0.03001), further emphasizing the distinct linguistic patterns in the harassment faced by mental health patients.

Lexical Differences. To further understand the linguistic differences between the online harassment messages received by mental health patients and the control group, we implement SAGE, a statistical model used for identifying distinguishing features in text data, particularly effective for uncovering the distinctive language patterns between different corpora (Eisenstein, Ahmed, and Xing 2011). SAGE identified the most salient words and phrases that differentiated the harassment messages received by the two groups, including both content words (e.g., specific insults or threats) and function words (e.g., pronouns, articles).

Table 3a shows the top SAGE words for the mental health patient group. Words like "dante" (SAGE = 3.941), "kellen" (SAGE = 2.946), and "huncho" (SAGE = 2.691) were among the most distinctive, alongside derogatory and offensive terms like "nigga" (SAGE = 2.185). This indicates a high presence of personal names and offensive language in the harassment directed at mental health patients, reflecting the personal and often abusive nature of these interactions.

In contrast, Table 3b highlights the top SAGE words for the healthy control group. Words such as "hugo" (SAGE = 5.090), "college" (SAGE = 1.901), and "ryan" (SAGE = 1.672) were prominent, alongside more neutral or positive words like "dating" (SAGE = 1.610), "classes" (SAGE = 1.585), and "hahahaha" (SAGE = 1.398). The presence of terms related to everyday activities and positive expressions suggests that the harassment faced by the healthy control group is less aggressive and personal.

Causal Relationship Between Adverse Mental Health Events and Harassment (RQ2)

Method

For RQ2, examining the impact of adverse mental health events on susceptibility to online harassment requires the use of causal methods, such that there is adequate mechanism to address the biases associated with the observed effects of adverse events (e.g., hospitalizations) on harassment.

While a randomized controlled trial (RCT) would provide a robust solution, it is neither feasible nor ethical in this context, given the sensitive nature of both harassment and mental health. Consequently, we employ an observational study design, leveraging methodologies and frameworks that have been used in prior social media research (Saha, Chandrasekharan, and De Choudhury 2019). Specifically, we implement a causal inference approach using matching, which seeks to approximate the conditions of an RCT by accounting for as many covariates as possible (Imbens and Rubin 2015). This approach is grounded in the potential outcomes framework, which evaluates whether a treatment T (mental health hospitalization) leads to an outcome (harassment) by comparing two possible outcomes: (1) $Y_i(T = 1)$ when T is applied, and (2) $Y_i(T = 0)$ when T is not applied. Using this framework, we structured a single-blinded experimental design described next.

Causal Inference Setup and Design. As each patient had at least one hospitalization event in our dataset, each hospitalization event was subject to an analysis that encompassed both pre-hospitalization (pre-HP) and post-hospitalization (post-HP) periods surrounding their hospitalization events. To investigate the association between adverse mental health events and online harassment, distinct time periods (n = $7, 14, 21, 28$ days) were established for each patient within the pre-HP and post-HP periods.

Understanding that online harassment experiences may trigger spillover effects across adjacent conversations, we

aggregated all messages within these time periods. This comprehensive approach accounted for scenarios where individuals withdrew from conversations immediately after encountering online harassment, providing a holistic view of their online behavior.

Recognizing that one's lived experiences influences how one perceives the world around them (Dredge, Gleeson, and De la Piedad Garcia 2014) and people would have their own baseline of mental health, we created *within-subject* control time periods for each patient participant. These control periods, which precede the participant's first observed hospitalization, effectively controlled for confounding factors that might arise due to differences between individuals. Each control period introduced placebo "hospitalization" events, corresponding to the hospitalization event of the participant. In addition, to control for the confounding factors that might influence the outcome, we controlled for the influence of seasonal effects on one's behavior or mood by assigning seasons to each message and used only those messages that fall within the same season of the corresponding hospitalization event. Each hospitalization event thus had an associated control group, allowing for a nuanced examination while mitigating the influence of confounding factors. Figure 1 provides an overview of the hospitalization and control periods for each participant.

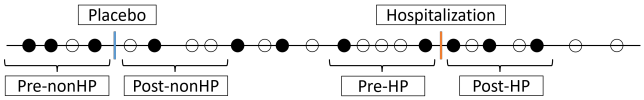


Figure 1: Timeline of the sent (white) and received (black) messages of patients diagnosed with mental illness with time windows surrounding the hospitalization event and a corresponding control time window.

Difference-in-Difference (DID) Analysis. Based on the above setup, to accurately depict the relationship between adverse events such as psychiatric hospitalization and the likelihood of receiving online harassment, we employed a Difference-in-Differences (DID) analysis – a widely used quasi-experimental research approach that existed from mid-19th century that 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). Utilizing the online harassment classifier from RQ1, we applied it to all received messages per patient or control participant. Then we extracted an average online harassment message frequency ratio or probability for each pre-HP ($p_b(g)$) and post-HP ($p_a(g)$) period, associated with each hospitalization event of a patient or control participant (g=patient, control). Then for each hospitalization event, we measured the difference, D_g, to represent the average difference of the harassment probabilities surrounding the hospitalization of a patient ($D_{g=\text{patient}}$) or a control participant ($D_{g=\text{control}}$):

$$D_g = p_a(g) - p_b(g) \quad (1)$$

Using the two differences $D_{g=\text{patient}}$ and $D_{g=\text{control}}$, we then measured the DID of the treatment group with respect to the control group–more specifically, $DID = D_{g=\text{patient}} - D_{g=\text{control}}$. This was performed for each of the four time windows: $n = 7, 14, 21, 28$ days.

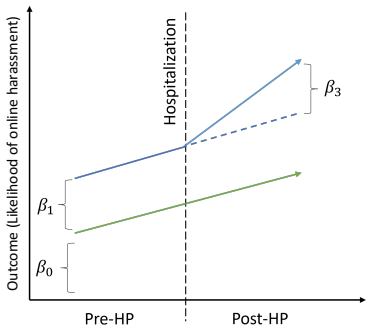


Figure 2: DID conceptual diagram showing coefficients used in the OLS regression model (eqn. 2).

Time Window	$D_{g=\text{patient}}$	$D_{g=\text{control}}$	DID
7 days	0.02996	0.00776	**0.02221**
14 days	0.05544	-0.00395	**0.05939**
21 days	0.05212	0.00413	**0.04799**
28 days	0.01521	0.00639	**0.00883**

Table 4: DID analysis comparing the changes in the probability of receiving online harassment messages before and after hospitalization across different time windows.

To ensure robustness of our findings from the DID analysis, we employed an approach from prior research (Bertrand, Duflo, and Mullainathan 2004) wherein we trained an Ordinary Least-Squares (OLS) Regression Model on $p_a(g)$ and $p_b(g)$ that are used to compare the patient and control groups. The OLS model used three variables to quantify the statistical significance of the DID analysis: q (whether it corresponded to the patient or the control groups), t (whether it corresponded to the pre-HP or post-HP period), and categorical variables as dummy variables. This OLS approach allowed us to establish that the observed difference (DID) between the feature changes in the treatment group and those in the control group did not occur out of random chance but rather due to the treatment, which in this case are the hospitalization events. The regression model is below:

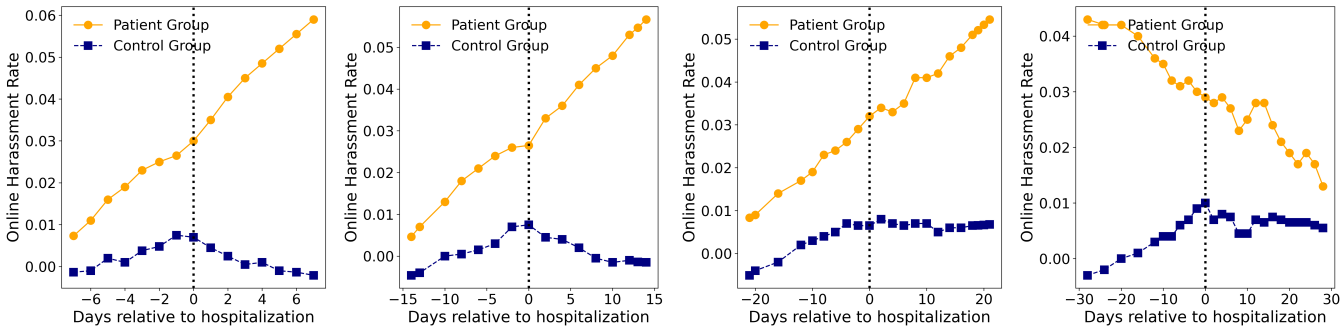
$$Y = \beta_0 + \beta_1 * q + \beta_2 * t + \beta_3 * (q * t) + \Sigma_{i=1}^{n} r_i \quad (2)$$

Figure 2 shows a conceptual representation of DID in relation to the OLS model.

In summary, by integrating DID analysis with OLS regression, we provide a comprehensive and statistically rigorous examination of the impact of hospitalization on online harassment among mental health patients. This combined approach contributes valuable insights to the understanding of online harassment dynamics in vulnerable populations.

Results

To recap, the difference-in-differences (DID) analysis was conducted to evaluate the impact of hospitalization on the



(a) 7 day time window (b) 14 day time window (c) 21 day time window (d) 28 day time window

Figure 3: Trends of the likelihood of online harassment during pre-HP and post-HP periods.

likelihood of receiving online harassment messages among mental health patients compared to a control group.

After testing for DID assumptions using the approach detailed in Appendix B, in Table 4, we present our major findings. The results in Table 4 demonstrate significant differences in the change in online harassment between the two groups ($(F\text{-stat: } 1.75^{30}, 6.88^{35}, 1.83^{32}, 6.76^{31}; (p < .05)$ and effect size of $\beta_3 : \eta^2 = 0.74, 0.82, 0.71, 0.67; (p < .05)$ for $n = 7, 14, 21, 28$, based on the OLS model in eqn. 2). For the 7-day time window, the patient group experienced an increase in harassment messages by 0.02996, while the control group saw a smaller increase of 0.00776, resulting in a DID estimate of 0.02221. This positive and statistically significant DID value indicates that mental health patients were more likely to receive online harassment messages post-hospitalization compared to the control group within the first week. In the 14-day time window, the difference became more pronounced. The patient group showed a substantial increase of 0.05544 in harassment messages, whereas the control group experienced a slight decrease of -0.00395. This resulted in a DID estimate of 0.05939, the highest observed in our analysis, suggesting a significant rise in harassment directed at mental health patients in the two weeks following hospitalization.

Further, for the 21-day time window, the patient group had an increase of 0.05212 in harassment messages, compared to a small increase of 0.00413 in the control group. The resulting DID estimate was 0.04799, indicating that the elevated harassment levels persisted for mental health patients three weeks post-hospitalization, albeit with a slightly reduced magnitude compared to the 14-day window. Finally, in the 28-day time window, the patient group experienced a modest increase of 0.01521 in harassment messages, while the control group saw a smaller increase of 0.00639. The DID estimate for this period was 0.00883, indicating that although the difference in harassment levels between the two groups remained positive, the effect diminished over the four weeks post-hospitalization.

Overall, as can also be observed in Figure 3, the DID analysis reveals that hospitalization is associated with a notable increase in the likelihood of receiving online harassment messages for mental health patients, particularly in the first two weeks following hospitalization. It is worth noting that the pattern in Figure 3d (28-day window) diverges from the shorter windows. This attenuation is expected: as the observation period lengthens, the acute post-hospitalization spike in harassment becomes diluted by subsequent weeks of relatively stable activity, yielding a weaker DID effect. In other words, while harassment risk is elevated most sharply in the immediate 1–2 weeks following hospitalization, the signal naturally diminishes over a longer 28-day span.

To further validate that the observed post-hospitalization spike in harassment is not merely a short-term fluctuation, we conducted a longitudinal analysis using linear mixed effects modeling. Results in Appendix C confirm both an immediate increase and a gradual decline in harassment over time, strengthening the above temporal interpretation.

These findings underscore the vulnerability of mental health patients to online harassment during critical periods of recovery and highlight the need for targeted interventions to mitigate this risk. The temporal pattern observed suggests that the impact is most acute in the immediate aftermath of hospitalization and gradually attenuates over time, though it remains significant up to four weeks post-hospitalization.

Can Social Support and Resilience Reduce the Risk of Harassment? (RQ3)

RQ3 seeks to explore if an individual's resilience and social support received online can reduce the likelihood of receiving online harassment following adverse events.

Method

To study this relationship, we adopted the econometric technique of Granger causality analysis (Geweke 1984). Granger causality is a statistical hypothesis test used to determine whether one time series can predict another time series. It is widely used in econometrics. Granger causality analysis rests on the assumption that if X causes Y then changes in X will systematically occur before changes in Y. For RQ3, the variable X will be the social support in the received messages or the resilience in the sent messages of the mental health patients, and Y will capture the likelihood of online harassment. Figure A2 illustrates this Granger framework.

Below we describe the predictor variables X: resilience and social support, used here.

Resilience Scores Shown in Figure A2a, one of the goals of our Granger causality framework is to assess whether resilience expressed in historical messages of a patient can forecast their corresponding trend of receiving harassment.

In the absence of ground truth data on resilience, we utilized Kang et al. (2022)'s study which aimed to construct a resilience dictionary to quantify one's resilience to crises and adverse events in online communities. The study implemented a systematic analysis pipeline that utilized posts from a mental health forum run by Schizophrenia: A National Emergency (SANE) Australia, discovered core themes, conceptualized resilience indicators, and generated a resilience dictionary. The complete resilience dictionary is comprised of 132 unique terms, each with 5 most similar terms for each descriptive term. Normalizing the count of these terms in each sent message of the patient, we assigned resilience scores to each message sent by a patient.

Furthermore, we drew upon prior literature that provided valuable insights into the different concept of resilience (McLaughlin et al. 2021). The study identified 1) age and stage and 2) trial and error resilience from the interviews. Age and stage resilience referred to the relationship between the maturity and the employed resilience strategies while trial and error resilience denoted the strategies that were employed and then abandoned if they did not seem to help with coping.

Using chain-of-thought based prompting, we then used examples from McLaughlin et al. (2021)'s study to prompt a GPT-4 large language model (LLM) to first learn the youth's perception of resilience to adverse events. Then, using the resilience dictionary above (Kang et al. 2022), we asked the LLM to generate a dataset of 330 resilience messages and 330 non-resilience messages. Table A1 provides this prompting strategy. Appendix D elaborates further on our approach.

Then we trained a BERT classifier on the LLM-generated dataset to create a binary resilience classifier. The classifier showed an accuracy of 0.91, precision of 0.92, recall of 0.90, F1-score of 0.91, and an Area under the ROC Curve (AUC) of 0.91. The classifier was applied to each sent message of the mental health patients to determine whether the message expressed resilience. Classifier reliability was evaluated using human annotations as described in Appendix E.

Social Support Scores Finally, per Figure A2b, with the Granger causality framework, we also intended to assess whether receiving support in Instagram conversations in the past could offset harassment, or forecast its trend. To quantify the social support in the messages received by the mental health patients, we adopted an informational support (IS) and emotional support (ES) scoring system from a prior study on support-seeking behaviors (Kim et al. 2023). The scoring system was founded upon the "Social Support Behavioral Code" (Sharma and De Choudhury 2018); which defines five types of support–informational, emotional, instrumental, esteem, and network support. For simplification, adapting the Likert scale classification theme (1=least supportive to 3=most supportive) from a study that used this framework (Sharma and De Choudhury 2018), the scoring system was built by developing a deep-learning model based on BERT, with ground truth scored by human annotators. The model achieved accuracy of 72.5%/71.3% and F1-scores of 0.726/0.715 for IS and ES, respectively. We applied the model to label ES and IS in each received message, providing insights into the recipients' support networks.

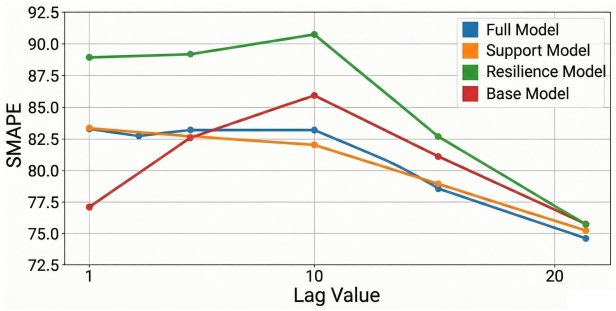


Figure 4: SMAPE of each VAR model.

Vector Autoregressive Modeling Approach Using guidance from prior work (De Choudhury, Kumar, and Weber 2017), we trained a total of four stepped Vector Autoregressive (VAR) models to assess the extent to which individual resilience or support received online can alleviate the experience of harassment. Since we expect the predictor variables here to be negatively associated with the outcome, we computed a complement of the outcome – in this case, it being the proportion of non-harassment messages received by a mental health patient. Our VAR models forecasted this non-harassment score on day t_k based on the predictor variables of n previous days $(t_{k-1}, t_{k-2}, ..., t_{k-n})$. We describe each VAR model as follows:

- **Base Model**: This model utilizes only linguistic features derived from a patient's Instagram messages, including categories from LIWC, activation and dominance categories from the Affective Norms for English Words (ANEW) lexicon (Bradley and Lang 1999), and the Flesch-Kincaid Grade Level readability metric.
- **Resilience Model**: In addition to the linguistic features in the Base Model, this model incorporated the resilience scores calculated on the sent messages of the patients.
- **Support Model**: This model extends the Base Model by including social support scores (ES and IS).
- **Full Model**: This comprehensive model integrated all the features used in the previous models: linguistic features, resilience scores, and social support scores, to provide the most holistic forecast of non-harassment scores.

Results

We begin by examining whether the pairs of time series we consider for the Granger causality test satisfy the Dickey Fuller tests for stationary time series (Dickey and Fuller 1981). This is an important step to ensure the Granger

causality tests are applicable to the different pairs of time series we consider here (e.g., the trend of resilience scores and the trend of non-harassment messages received). We find that the pairs satisfy the Dickey Fuller test ($p < .05$).

Model Performance Evaluation Per Figure 4, we observe the trend of Symmetric Mean Absolute Percentage Error or SMAPE for the Base, Support, Resilience, and Full models across different lag values. The Base model, which included only linguistic features, exhibited relatively high SMAPE values, indicating lower predictive accuracy. The addition of informational and emotional support scores in the Support model improved the forecasting performance, as evidenced by lower SMAPE values compared to the Base model. The Resilience model, which incorporated resilience scores, showed varying performance, generally worse than the Support model but better than the Base model at specific, especially lower lag values. The Full model, integrating all features (linguistic, support, and resilience scores), consistently demonstrated the best performance, achieving the lowest SMAPE values across most lag values.

Overall this indicates that a patient's resilience, as expressed in Instagram conversations, and the support they receive on that platform, tend to bear at least some modest explanatory power against harassment trends, above and beyond what could be explained with their linguistic patterns alone. This is especially true for lower lag values, that is, when the resilience and support are observed in recent history ($\sim$7 days). Support tends to offset the risk of forthcoming harassment to a greater degree ($\sim$22%) over resilience, indicating that interpersonal relationships play a key role in combating negative experiences online. Upon combining all of these variables together, the fact that the Full model gives the lowest SMAPE indicates that approaches to tackle online harassment need to not only equip mental health patients with resiliency strategies, but also provide them avenues to inculcate nurturing social relationships online.

Predictor Significance and Qualitative Insights Observing Rao's approximate F-statistics and p-values (based on Canonical-Correlation Analysis between the two sets of time series, for each pair of predictor and predicted variables in the VAR models) further demonstrates the unique roles played by resilience and social support in mitigating the risk of harassment. See Figure A3 for key terms appearing in Instagram messages of patients, demonstrating having received informational and emotional support and those expressing the author's resiliency. Appendix G unpacks additional qualitative insights around how support and resilience language surfaces in online harassment conversations.

In the Full Model, the informational support score emerged as the most significant predictor with an extremely high F-statistic (1.59E+03), a low Wilks' λ (likelihood ratio) statistic of 0.17, and a significant p-value $< 10^{-15}$, showing it could act as a protective factor by giving patients the tools they need to avoid or mitigate the effects of harassment, creating a sense of control over their interactions. It may also highlight the importance of education and practical advice in reducing vulnerability to online harassment. Aside from the high level overview in Figure A3a, note the following paraphrased message received by a mental health patients, that scored high on informational support:

> "Bro, that sucks, I'm sorry. You ever try that 5-4-3-2-1 thing? It's like, you name 5 things you can see, 4 you can touch, 3 you can hear, 2 you can smell, and 1 you can taste. It sounds kinda weird but it helps me chill when I'm freaking out. Oh, and deep breaths! Like, breathe in for 4 seconds, hold for 4, and breathe out for 4. Might not fix everything, but it helps me calm down sometimes."

Next, the emotional support score with an F-statistic of 2.79E+02, a Wilks' λ statistic of 0.29, and a p-value of $< 10^{-15}$ further indicates how it also acts as a protective buffer against harassment by strengthening the individual's emotional resources, reinforcing their support network, and enhancing their ability to cope with and respond to negative behaviors. Complementing our prior observations, it showcases the importance of empathy, care, and social connections in reducing vulnerability to harassment and promoting safer online environments – an insight that we can glean from Figure A3b and the following paraphrased message:

> "Ugh, I'm sorry you're going through that. That sucks, I know how rough it can get. Just want you to know I'm here, no matter what. You don't have to deal with this alone."

Finally, although to a lesser extent, resilience scores had a F-statistic of 1.08E+02 and Wilks' λ statistic of 0.36 ($p < 10^{-10}$). This may indicate that expressing resilience in conversations can reduce the risk of harassment by signaling emotional strength, non-reactivity, and assertiveness. It may reveal that the person is not easily manipulated, making them a less attractive target to harassers. It could also communicate a refusal to engage with negativity and a focus on overcoming challenges, which can ultimately discourage hostile interactions. There were many expressions of resiliency in the messages shared by the patients (ref. Figure A3c); e.g., the one below wherein the author shows a determination to stop tolerating disrespect, indicating a shift in strategy based on past experiences:

> "[...] like no I am done getting disrespected. My abuser can just walk away"

Discussion

Understanding the Vulnerabilities of Mental Health Patients to Online Harassment

Our results corroborate previous findings which have identified individuals with mental health challenges as more vulnerable to online harassment Dredge, Gleeson, and De la Piedad Garcia (2014). The linguistic analysis in this study identified distinct patterns of harassment messages directed at mental health patients, such as a higher prevalence of second-person pronouns, swear words, and personal names. Importantly, harassment messages directed at mental health patients often lack tentativeness or auxiliary verbs, suggesting that harassers may adopt a more assertive and direct communication style. Broadly, this aligns with previous research that has documented the aggressive, personal, and offensive nature of such abuse (Ozden and Icellioglu 2014). Specifically, the increased use of derogatory language and

personalized insults observed in the harassment directed at mental health patients mirrors the findings of Gualdo et al. (2015), who similarly noted that individuals experiencing mental health crises are frequently subjected to more hostile online interactions.

This linguistic differentiation has practical implications for harassment detection systems. Building on previous work by Razi et al. (2022), which emphasized the importance of victim-labeled harassment data and consideration of conversational context, our study suggests that future harassment detection models could benefit from incorporating the unique linguistic patterns underlying harassment messages, such as tone and context, to identify instances particularly targeting vulnerable populations. Beyond supporting moderation efforts, the distinctive linguistic patterns of harassment targeting individuals with mental health struggles raise important questions about the cyclical relationship between mental health and online abuse (Kim et al. 2024). Our results show that psychiatric hospitalization—a clear marker of acute distress – often precedes a spike in harassment, particularly in the two weeks post-discharge. This aligns with research suggesting that those with compromised mental health are seen as more vulnerable and may be more likely to attract abuse (Carmen, Rieker, and Mills 1984). Individuals' mental states shape their perception of social feedback, making this especially relevant in the context of social media, where many now seek support (Anderson and Jiang 2018).

Interventions to Mitigate the Impact of Harassment around Adverse Mental Health Events

Our findings reveal that psychiatric hospitalization marks a distinct period of vulnerability to online harassment, particularly in the two weeks immediately following discharge. For researchers and clinicians, consistent with guidance in the literature (Layne et al. 2008), the identification of post-hospitalization periods as high-risk windows – as also demonstrated by Ernala et al. (2022) – can inform the timing of digital mental health interventions. This can include targeted outreach or monitoring strategies. Computational models could be designed to detect signals of post-crisis vulnerability and deploy automated yet supportive nudges, such as promoting resilience, building content or directing users to peer support resources (Sharma and De Choudhury 2018; Kim et al. 2023). For platform designers, our results suggest the need for temporally adaptive content moderation: e.g., implementing heightened harassment detection sensitivity or protective friction (e.g., comment delays or warning prompts) during these critical windows. Additionally, platforms could allow users to voluntarily flag recent hospitalizations or high-stress periods – privately and securely – to receive tailored protections without compromising privacy.

It is also important to situate these findings in the contemporary digital landscape. Since the time of our data collection (2016–2020), the rise of generative AI has transformed how both content and moderation operate online. AI-mediated communication may amplify harassment risks through synthetic text, deepfake imagery, or large-scale automated trolling that were not present in our dataset. At the same time, AI-based moderation tools now increasingly determine which harmful messages are surfaced, flagged, or suppressed, shaping user experiences in ways that differ from earlier platform environments. These shifts underscore that while the psychosocial mechanisms identified here remain salient, their empirical expression may evolve, calling for replication in AI-mediated settings.

Finally, for policymakers, our findings support the case for trauma-informed design standards in digital mental health ecosystems (Chen et al. 2022). Guidelines should encourage platforms to adopt proactive safeguarding mechanisms that respond to dynamic user contexts—not just static account behaviors. Together, these actionable steps can translate our longitudinal evidence into interventions that are timely, human-centered, and clinically informed.

Social Support and Resilience: Protective Role Against Harassment

A key contribution of this study is the examination of how social support and personal resilience mediate the impact of online harassment. Our findings suggest that individuals who receive higher levels of emotional and informational support are less likely to experience harassment. This aligns with prior work showing that strong online networks can buffer stressors and improve outcomes for vulnerable populations (Sharma and De Choudhury 2018; Kim et al. 2023). Informational support, in particular, stood out in our analysis – messages offering coping strategies or grounding techniques were associated with reduced harassment exposure, likely by empowering individuals to manage interpersonal conflicts more effectively.

These findings point to opportunities for preventive intervention. Platforms could promote support-rich exchanges during periods of risk by amplifying helpful messages through in-app prompts or recommendations. Moderation tools might be adapted to flag conversations where informational or emotional support is lacking and suggest low-friction support resources. Additionally, building peer-based support communities or resilience training features – especially for those with known histories of crisis – may offer preemptive protection against future harassment escalations.

Although resilience played a more modest role than support, it remains significant. Expressions of boundary-setting or non-reactivity were correlated with fewer hostile responses, suggesting that resilience not only reflects coping but may actively deter aggressors (McLaughlin et al. 2021). However, this must be balanced with prior work noting that overt resistance or emotional reactivity can sometimes provoke further harassment in adversarial contexts (Cook, Schaafsma, and Antheunis 2018). Future interventions could therefore focus on fostering communication skills that blend strength with de-escalation, minimizing retaliation risks. Together, these insights underscore the dual importance of social and individual-level factors in both responding to and preventing harassment in vulnerable digital contexts.

At the same time, resilience is not uniformly protective. Certain coping strategies—such as boundary-setting, assertive refusal, or non-reactivity—can signal strength and help deter harassment, but in adversarial environments they may also provoke retaliation or escalate hostility. This du-

ality has been noted in prior literature and highlights that resilience is highly context-dependent. Interventions should therefore not only encourage resilient communication but also equip individuals with strategies for de-escalation and safe boundary management, ensuring that resilience functions as a buffer rather than a trigger for further harm.

Limitations, Conclusions, and Future Work

This study provides novel insights into the relationship between adverse mental health events and online harassment. We show that individuals are particularly vulnerable to harassment during critical periods, such as immediately after psychiatric hospitalization, and that social support and resilience can mitigate this risk. However, several limitations should be noted. First, the dataset—Instagram direct messages from a specific cohort—may not generalize across platforms, populations, or evolving digital ecosystems. It also spans 2016–2020, predating COVID-19 and the widespread integration of generative AI. While we argue that the underlying psychosocial mechanisms (e.g., post-hospitalization vulnerability, protective effects of resilience) remain conceptually robust, their manifestation may differ in today's AI-mediated environments (see Appendix A).

Methodological considerations additionally warrant caution. Some classifiers were trained or validated on partially LLM-generated data, which, while addressing sparsity, may introduce artificial patterns. This is particularly relevant for constructs like resilience and social support, where synthetic data may oversimplify real-world heterogeneity. These limitations also raise ethical concerns: mischaracterizing such signals could lead to inappropriate interpretations or interventions. Future work should prioritize participant-annotated datasets and participatory approaches to better capture lived experience and improve both validity and ethical robustness.

Ethical Statement

Our work examines sensitive data from a vulnerable population—youth with mental health challenges—and thus prioritizes strong ethical safeguards. The study was approved by Institutional Review Boards and adhered to strict privacy protections, including deidentification, paraphrasing of content, and reliance on NIH Certificates of Confidentiality to minimize re-identification risk emphasized as a risk in prior research (Ayers et al. 2018; Nguyen et al. 2022). Data access followed a principle of least privilege and was secured on HIPAA-compliant infrastructure, with involvement from clinical collaborators to ensure trauma-informed interpretation (Chen et al. 2022). At the same time, we recognize potential risks: findings could unintentionally reinforce stigma, enable victim-blaming, or discourage online participation among already vulnerable individuals. Additionally, computational tools for harassment detection may introduce privacy concerns or misuse if deployed without safeguards. We therefore emphasize that all applications of this work should adopt a trauma-informed, patient-centered lens, ensuring that insights are used to protect and empower individuals rather than restrict or surveil them. Additional ethical considerations are provided in the Paper Checklist. Authors declare no competing interests.

Acknowledgments

Authors were partly supported through National Institute of Mental Health grant R01MH117172.

References

Ali, S.; Razi, A.; Kim, S.; Alsoubai, A.; Ling, C.; De Choudhury, M.; Wisniewski, P.; and Stringhini, G. 2023. Getting meta: A multimodal approach for detecting unsafe conversations within instagram direct messages of youth. *Proc ACM-HCI*, 7(CSCW1): 1–30.

Aljasir, S.; and Alsebaei, M. 2022. Cyberbullying and cybervictimization on digital media platforms: the role of demographic variables and parental mediation strategies. *Humanities and Social Sciences Communications*, 9(1): 1–9.

Anderson, M.; and Jiang, J. 2018. Teens, Social Media & Technology 2018 | Pew Research Center.

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

Arshed, M. A.; Samreen, Z.; Ahmad, A.; Amjad, L.; Muavia, H.; Dewi, C.; and Kabir, M. 2025. Multi-Class Visual Cyberbullying Detection Using Deep Neural Networks and the CVID Dataset. *Information*, 16(8): 630.

Ayers, J. W.; Caputi, T. L.; Nebeker, C.; and Dredze, M. 2018. Don't quote me: reverse identification of research participants in social media studies. *NPJ digital medicine*, 1(1): 1–2.

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.

Bradley, M. M.; and Lang, P. J. 1999. Affective norms for English words (ANEW): Instruction manual and affective ratings. Technical report, The center for research in psychophysiology.

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

Campbell, M. A. 2005. Cyber bullying: An old problem in a new guise? *J. Psychologists & Counsellors in Schools*, 15(1): 68–76.

Carmen, E.; Rieker, P. P.; and Mills, T. 1984. Victims of violence and psychiatric illness. In *The gender gap in psychotherapy: Social realities and psychological processes*, 199–211. Springer.

Chen, J. X.; McDonald, A.; Zou, Y.; Tseng, E.; Roundy, K. A.; Tamersoy, A.; Schaub, F.; Ristenpart, T.; and Dell, N. 2022. Trauma-informed computing: Towards safer technology experiences for all. In *Proc CHI*, 1–20.

Conneau, A.; Kiela, D.; Schwenk, H.; Barrault, L.; and Bordes, A. 2017. Supervised learning of universal sentence representations from natural language inference data. *arXiv preprint arXiv:1705.02364*.

Cook, C.; Schaafsma, J.; and Antheunis, M. 2018. Under the bridge: An in-depth examination of online trolling in the gaming context. *New Media & Society*, 20(9): 3323–3340.

De Choudhury, M.; Kumar, M.; and Weber, I. 2017. Computational approaches toward integrating quantified self sensing and social media. In *Proc CSCW*, 1334–1349.

Dickey, D.; and Fuller, W. 1981. Likelihood ratio statistics for autoregressive time series with a unit root. *Econometrica: journal of the Econometric Society*, 1057–1072.

Dredge, R.; Gleeson, J.; and De la Piedad Garcia, X. 2014. Cyberbullying in social networking sites: An adolescent victim's perspective. *Computers in human behavior*, 36: 13–20.

Eisenstein, J.; Ahmed, A.; and Xing, E. P. 2011. Sparse additive generative models of text. In *Proc ICML*, 1041–1048.

Ernala, S. K.; Seybolt, J.; Yoo, D. W.; Birnbaum, M. L.; Kane, J. M.; and De Choudhury, M. 2022. The reintegration journey following a psychiatric hospitalization: examining the role of social technologies. *Proc ACM-HCI*, 6(CSCW1): 1–31.

Eroğlu, Y.; Aktepe, E.; Akbaba, S.; Işık, A.; and Özkorumak, E. 2015. Investigation of prevalence and risk factors associated with cyberbullying & victimization. *Education & Science*, 40(177).

Gashroo, O. B.; and Mehrotra, M. 2024. DetectHATE: Detecting Targeted Hate-A Framework for Classifying Online Abuse on X. *International Journal of Performability Engineering*, 20(11).

Geweke, J. F. 1984. Measures of conditional linear dependence and feedback between time series. *JASA*, 79(388): 907–915.

Gualdo, A.; Hunter, S.; Durkin, K.; Arnaiz, P.; and Maquilón, J. 2015. Emotional impact of cyberbullying: Differences in perceptions & experiences as a function of role. *Comp & Ed*, 82: 228–235.

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

John, A.; Glendenning, A.; Marchant, A.; Montgomery, P.; Stewart, A.; Wood, S.; Lloyd, K.; et al. 2018. Self-harm, suicidal behaviours, and cyberbullying in children and young people: Systematic review. *JMIR*, 20(4): e9044.

Kang, Y.-B.; McCosker, A.; Kamstra, P.; and Farmer, J. 2022. Resilience in web-based mental health communities: Building a resilience dictionary with semiautomatic text analysis. *JMIR Formative Research*, 6(9): e39013.

Kim, M.; Saha, K.; De Choudhury, M.; and Choi, D. 2023. Supporters First: Understanding Online Social Support on Mental Health from a Supporter Perspective. *Proc. ACM-HCI*, 7(CSCW1): 1–28.

Kim, S.; Razi, A.; Alsoubai, A.; Wisniewski, P. J.; and De Choudhury, M. 2024. Assessing the Impact of Online Harassment on Youth Mental Health in Private Networked Spaces. In *Proc ICWSM*, volume 18, 826–838.

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

Layne, C. M.; Beck, C. J.; Rimmasch, H.; Southwick, J. S.; Moreno, M. A.; and Hobfoll, S. E. 2008. Promoting "resilient" posttraumatic adjustment in childhood and beyond:"Unpacking" life events, adjustment trajectories, resources, and interventions. In *Treating traumatized children*, 31–66. Routledge.

Lenhart, A.; Smith, A.; Anderson, M.; Duggan, M.; and Perrin, A. 2015. Teens, technology and friendships.

Madhurima, S.; Ajith, K. A.; Swathy, V.; and Prathap, B. R. 2024. Abusive Words Detection on Reddit Comments Using Machine Learning Algorithms. In *2024 DICCT*, 312–317. IEEE.

Marshan, A.; Nizar, F.; Ioannou, A.; and Spanaki, K. 2025. Comparing machine learning and deep learning techniques for text analytics: Detecting the severity of hate comments online. *Information Systems Frontiers*, 27(2): 487–505.

McLaughlin, P.; Kennedy, B.; Harris, A.; Hamilton, M.; Richardson, J.; and Holman-Jones, S. 2021. Online and social media resilience in young people in vulnerable contexts. *Vulnerable children and youth studies*, 16(2): 178–188.

Nguyen, V. C.; Lu, N.; Kane, J. M.; Birnbaum, M. L.; and De Choudhury, M. 2022. Cross-Platform Detection of Psychiatric Hospitalization via Social Media Data: Comparison Study. *JMIR Mental Health*, 9(12): e39747.

Ozden, M.; and Icellioglu, S. 2014. Perception of cyberbullying and cybervictimization by university students in terms of personality factors. *Procedia-SBS*, 116: 4379–4383.

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 CHI-Extended Abstracts*.

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

Ronis, S.; and Slaunwhite, A. 2019. Gender and geographic predictors of cyberbullying victimization, perpetration, & coping modalities among youth. *Canadian J. of school psych*, 34(1): 3–21.

Saha, K.; Chandrasekharan, E.; and De Choudhury, M. 2019. Prevalence and psychological effects of hateful speech in online college communities. In *Proc. WebSci*, 255–264.

Sharma, E.; and De Choudhury, M. 2018. Mental health support and its relationship to linguistic accommodation in online communities. In *Proc CHI*, 1–13.

Todorovic, M.; Kozakijevic, S.; Jovanovic, L.; Babic, L.; et al. 2025. Detecting Harassment in User Comments: 2-Tier Machine Learning & Metaheuristics Approach with Natural Language Processing. *SNCS*, 6(6): 573.

Ybarra, M.; and Mitchell, K. 2007. Prevalence and frequency of Internet harassment instigation: Implications for adolescent health. *J. Adolescent Health*, 41(2): 189–195.

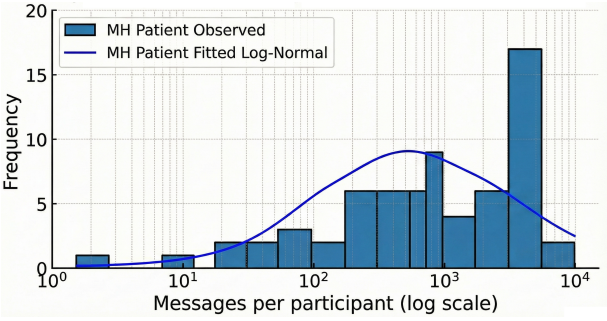
Zimmerman, M.; and Mattia, J. I. 2001. A self-report scale to help make psychiatric diagnoses: the Psychiatric Diagnostic Screening Questionnaire. *Archives of general psychiatry*, 58(8): 787–794.

AAAI ICWSM Paper Checklist

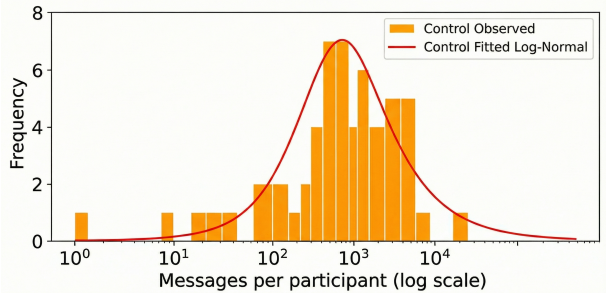
1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes, see the Data section and Ethics Statement above.
   (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 section.
   (e) Did you describe the limitations of your work? Yes, see Limitations, Conclusion, and Future Work.
   (f) Did you discuss any potential negative societal impacts of your work? Yes we clearly describe the implications of our work in the Discussion section, including negative implications in the Ethics Statement.
   (g) Did you discuss any potential misuse of your work? Yes, in the Ethics Statement.
   (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 section and the Ethics Statement.
   (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)? No, we do not include the data or the trained models used in this work because the guiding IRB approval does not allow resharing of the sensitive social media and clinical data of participants. Moreover, the consent process we utilized in this data collection (ref. Data) only permits using the data by the authorized research team. Next, trained machine learning models are not released here due to risk of misuse, or risk of potential traceability to the participants of the study. That said, we have included detailed information about both the data and our machine learning models which should make them easily replicable and reproducible in other research contexts.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes, in the "Detecting Online Harassment Messages" subsection of RQ1.
   (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)? Yes, in the "Method" subsection of RQ3.
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes, we do that individually for each of the three RQs; for instance, within "Linguistic Comparison" for RQ1, within "Results" for RQ2, and in "Model Performance Evaluation" for RQ3.
   (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 the Data section (Social Media Data, Clinical Data) and "Detecting Online Harassment Messages" in the RQ1 section.
   (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? Yes, in the Data section.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes, in the Ethics Statement.
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? NA
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? 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? This is not applicable because the authors of this paper appropriated datasets collected by others in unrelated projects that have been cited throughout the paper.
   (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? Not applicable; see above.
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? Not applicable; see above.
   (d) Did you discuss how data is stored, shared, and deidentified? Yes, in the Ethics Statement.

Appendix



(a) Patient participants



(b) Control participants

Figure A1: Message distributions across participants for mental health (MH) patient and control groups.

A. Generalizability, Scope, and Representativeness of the Instagram Dataset

Additional details on recruitment and clinical assessments. For patients, diagnostic confirmation was based not only on medical record review but also on structured clinical interviews conducted at the time of their most recent hospitalization, with diagnoses cross-verified by at least two treating clinicians. This dual confirmation reduced the likelihood of misclassification due to documentation errors or transient symptom presentations. For controls, in addition to the SCID or Psychiatric Diagnostic Screening Questionnaire, participants were required to have no record of psychotropic medication use in the prior 12 months, a safeguard not detailed in the main text. All recruitment sites maintained a weekly monitoring process to ensure that any new psychiatric diagnoses emerging during the study window were flagged, and those individuals were excluded from the control pool.

Platform and Generalizability Considerations. A central feature of this study is its focus on Instagram, which was selected due to its widespread adoption among youth and young adults – the demographic most at risk for both mental health distress and online harassment. According to Pew Research data, Instagram remains one of the most actively used platforms among teens and young adults in the United States (Anderson and Jiang 2018), particularly for private and semi-private communication like direct messages. As such, Instagram serves as a highly ecologically valid context for studying interpersonal online dynamics during critical mental health events. While findings may not generalize directly to all platforms (e.g., Reddit, TikTok), the conversational nature and visual emphasis of Instagram offer key insights into how harassment and support unfold in youth digital communication. Moreover, as messaging functionality becomes increasingly central across platforms, the patterns identified here may be indicative of broader trends in private online interactions elsewhere.

Furthermore, although our analysis was conducted over 134 participants, which may appear limited in size, the dataset encompasses over 360,000 private Instagram direct messages – allowing for a longitudinal, person-centered analysis of behavioral patterns surrounding clinically validated adverse mental health events. Our analytical focus is not on cross-sectional generalization, but on capturing temporal dynamics within individuals over time. This level of data depth and ecological validity is rare, particularly in studies involving adverse health events like psychiatric hospitalization (Nguyen et al. 2022). Importantly, participant recruitment for such a cohort (mental health patients with DSM-5 diagnosed mental health conditions) is inherently time-consuming and resource-intensive, given the need for clinical verification, ethical safeguards, and the challenges of engaging a stigmatized and vulnerable population from a safety as well as logistical perspective. Our dataset thus reflects a deliberate methodological commitment to depth, context, and clinical rigor over breadth, consistent with prior work in computational social science and digital health (Ernala et al. 2022; Razi et al. 2022).

Dataset Balance and Representation. This dataset includes individuals clinically diagnosed with mental illness and verified healthy controls. In our filtered sample of 134 participants with valid Instagram data, the groups are balanced (50% each). However, patients contributed a higher proportion of the 360,000 messages. This disparity reflects natural variations in engagement patterns and the clinical objectives of the parent study, rather than systematic bias. To mitigate any potential analytical bias introduced by unequal

message volume, our core analyses, such as Difference-in-Differences and Vector Autoregressive modeling, were designed using within-subject comparisons and between-group controls. As a result, while the dataset may have limits in broader generalizability, it offers a rare and ecologically valid foundation for longitudinal, person-centered analysis of mental health and online interactions.

Reflections on Temporal Context and Current Digital Landscape. Our dataset spans 2016–2020, a period that predates the COVID-19 pandemic and the widespread integration of generative AI tools into online communication. While these temporal boundaries ensure a longitudinal and ecologically valid capture of youth social media interactions, it is important to recognize that today's digital landscape differs in several respects. First, the normalization of AI-mediated communication (e.g., chatbots, AI-generated images, text completions) may alter how harassment manifests, potentially introducing new modalities of abuse (e.g., deepfake harassment, automated trolling) that were not observable in our dataset. Second, platform policies and moderation practices have evolved considerably in recent years, with less algorithmic filtering and content warnings, but more commonly implemented user- or community-reporting systems adopted by social media platforms that may shift both the prevalence and visibility of harassment. Finally, social norms surrounding disclosure and support-seeking online have changed, with greater awareness of mental health but also heightened risks of context collapse and amplification through recommendation systems.

Summarily, these changes suggest that while our findings regarding the vulnerability of youth around hospitalization events and the protective role of support and resilience remain conceptually robust, their empirical expression in today's digital environments may differ. Future research could replicate these analyses on more recent, multi-platform datasets, explicitly considering AI-mediated interactions and contemporary moderation policies.

B. Testing for DID Assumptions

The DID approach comes with some assumptions, such as exchangeability (assumption of equivalent distribution outside of the treatment effect), positivity (assessing causal effects in subjects who are eligible for all levels of exposure we care about), and ruling out spillovers such as any confounding indirect impact on the outcomes of interest. The most important assumption to satisfy is, however, Stable Unit Treatment Value Assumption (SUTVA) that states that each unit's outcomes are independent of other units' treatment status (Imbens and Rubin 2015). In our case satisfying SUTVA would mean that we ensure that a participant's experience of online harassment does not depend upon whether the others in the dataset have had a psychiatric hospitalization. This is not the case for us, because the participants recruitment did not use any snowballing approach and each participant was recruited independently. Therefore, they are unlikely to be socially connected which might be otherwise responsible for such spillover effects of others' hospitalization on a different person's online harassment experience.

Next, unique to the DID approach, DID requires that in the absence of treatment, the (unobserved) difference between the treatment and control group, in terms of the outcomes of interest, is constant over time (Angrist and Pischke 2009). We checked for that by first building an autoregressive model (e.g., ARIMA) that projected the trends of the outcome variable of interest (probability of online harassment) in the treatment group into the post-hospitalization period, based on data in the pre-hospitalization period. Then, looking at the actual trend of the outcomes in the control group, and comparing that with the actual and ARIMA projected trends in the treatment group across both the pre- and post-hospitalization periods, we assessed their mutual difference. We found this difference to be consistent over time; a permanent difference (parallel trend or constant bias) helps to satisfy this DID assumption. For robustness, we performed this comparison of trends over different pre-hospitalization periods, as recommended in the literature (Imbens and Rubin 2015).

C: Longitudinal Modeling of Post-Hospitalization Harassment

To provide stronger empirical support for our claim that online harassment peaks following psychiatric hospitalization (RQ2), we performed a longitudinal statistical analysis using a linear mixed effects (LME) model. This approach allows us to model harassment trajectories over time while accounting for within-subject correlations in repeated observations.

Model Specification. We constructed a daily time series for each patient encompassing a symmetric window of 28 days before and after each hospitalization event. For each day t, we computed the proportion of received messages classified as harassment. Let y_{it} represent this score for individual i on day t. The model was specified as:

$$y_{it} = \beta_0 + \beta_1 \cdot \text{time}_{it} + \beta_2 \cdot \text{post}_{it} + \beta_3 \cdot (\text{time}_{it} \times \text{post}_{it}) + u_i + \epsilon_{it}$$

Here, time_{it} denotes the number of days since hospitalization (centered around zero), post_{it} is a binary indicator denoting the post-hospitalization period, and u_i is a participant-level random intercept. This model enables us to detect both an immediate shift in harassment levels post-hospitalization (β_2) and a change in trajectory (β_3).

Results. The model revealed a statistically significant level increase in harassment immediately following hospitalization ($\beta_2 = 0.031$, $p < 0.001$), confirming the sharp rise observed in our Difference-in-Differences analysis. The interaction term ($\beta_3 = -0.0014$, $p = 0.022$) was negative and statistically significant, indicating that while harassment peaks post-hospitalization, it gradually declines over time.

These findings suggest that hospitalization not only marks a vulnerable inflection point but also that the heightened harassment is not uniformly sustained. Instead, the risk is most acute in the immediate aftermath, underscoring the need for targeted platform-level support during this critical recovery window. Broadly, this longitudinal modeling approach strengthens our interpretation of harassment trends by moving beyond pre/post averages to assess continuous change in

response to RQ2. Importantly, it provides empirical grounding for temporal sensitivity in online safety interventions, reinforcing the argument we provide in the Discussion section for dynamic, trauma-informed moderation strategies that respond to life events such as psychiatric hospitalization.

Additional robustness checks. Beyond parallel trends tests, we examined whether harassment levels exhibited differential pre-hospitalization slopes between patients and controls by fitting placebo DID models on random "pseudo-hospitalization" dates at least 6 months prior to observed events. These placebo tests consistently yielded null results, supporting the robustness of our main DID estimates. Additionally, in the mixed-effects models, we tested for individual-level heterogeneity by including random slopes for time since hospitalization. This revealed that while the overall pattern of a post-hospitalization spike followed by gradual attenuation held, the magnitude of the spike varied considerably across patients, suggesting individualized vulnerability trajectories not captured in the aggregate analysis.

D. LLM-Based Resilience Scoring Strategy

You are tasked with creating a training set of social media messages that contain language reflecting personal and psychological resilience, particularly in the context of overcoming challenges or setbacks. Resilience is defined as the ability to recover quickly from difficulties, demonstrating mental toughness, optimism, and problem-solving. The context is online conversations from individuals who may face adversity, such as mental health challenges. Your goal is to generate exemplar but realistic messages shared by young people in Instagram. For this, use the attached resilience dictionary. Each message you produce should incorporate at least one or more of the resilience terms from the dictionary. You should use the following guidelines in the generation process. - Use conversational language that feels natural for someone sharing their thoughts or experiences on social platforms like Instagram - Each message should be encouraging, demonstrating emotional strength, and a proactive approach to handling stress or adversity - Ensure that the messages are framed positively, showing how the individual is coping or planning to overcome the difficulty Some example messages you can use for inspiration are given below. You should assume you are a youth who uses Instagram direct messages as the key source of communication with people. message 1 message 2 message 3 message 4 message 5

Table A1: Resilience message generation guidelines given to a GPT4 model, for the purpose of training a BERT classifier.

Theoretical Grounding of the Resilience Scoring Approach. Our resilience scoring approach relied on prior social science research that provides theoretical grounding as well as resources indicative of resilience in interpersonal interactions. We utilized two pieces of work as below.

First was the work of Kang et al. (2022) who implemented a systematic analysis pipeline that utilized posts from a mental health forum run by Schizophrenia: A National Emergency (SANE) Australia. These authors discovered core themes, conceptualized resilience indicators, and generated a resilience dictionary. We harnessed these dictionary terms as seeds to our subsequent approach.

Second, we utilized McLaughlin et al. (2021)'s study that examined young people aged 12-19 years in vulnerable contexts, interviewing six participants to explore themes of exposure and resilience. The study identified 1) age and stage and 2) trial and error resilience from the interviews. Age and stage resilience referred to the relationship between the maturity and the employed resilience strategies, while trial and error resilience denoted the strategies that were employed and then abandoned if they did not seem to help with coping. Examples of the two types of resilience are as follows: a) *Age and stage resilience:* "It took a fair few years to just wrap my head around it. Like I was getting down, like depressed . . . (unclear) and then just stop it. When I got older I got out of it and then like, not older, but also stopped and like just changed my whole group." b) *Trial and error resilience:* "I guess I just decided to take control myself- like I thought before I wasn't strong enough, but now I thought well I am- things just got better when I was in control."

Together, the dictionary of resilience terms and the framework of operationalizing resilience informed our chain of thought based prompt described in Table A1.

Assessing Quality of LLM Generations. To ensure the ecological validity and conceptual clarity of the LLM-generated dataset used to train the resilience classifier, we conducted a manual review of a stratified sample of 100 messages (50 resilience, 50 non-resilience). Two domain-informed annotators versed in youth digital communication and psychosocial resilience independently assessed if each message reflected its assigned label. Grounded in frameworks of psychological resilience and coping (McLaughlin et al. 2021), the annotators judged 91% of the messages as correctly labeled, with a Cohen's κ of 0.76 indicating strong agreement. Notably, resilience messages often featured boundary-setting, future-oriented thinking, or assertive self-reflection, while non-resilience messages leaned toward emotional reactivity or helplessness. These findings underscore the appropriateness of the LLM-generated messages as a training proxy and affirm their alignment with theoretically grounded constructs of resilience.

E. Evaluation of Resilience Classifier

To assess the robustness and reliability of the resilience scoring classifier developed for our study, we adopted two steps:

First, for the resilience classifier, we added a second-layer dropout regularization ($p = 0.4$) to reduce overfitting to synthetic LLM-generated samples, and we conducted an abla-

tion study where resilience scores were computed using only the Kang et al. (2022) terms without LLM augmentation. Results showed weaker predictive performance (F1=0.78 vs. 0.91), reinforcing the utility of the augmented approach but also highlighting its dependence on synthetic data.

Next, we undertook a human-centered validation procedure informed by established qualitative and mixed-methods research practices in social computing. Recall that the classifier was designed to detect resilience in messages authored by mental health patients, drawing on a curated dictionary of resilience terms (Kang et al. 2022) and fine-tuned using synthetically generated messages guided by real-world examples (McLaughlin et al. 2021). Yet, given the sensitive context of mental health and the complexity of subjective constructs like resilience, it was essential to ensure that the model's predictions aligned with human judgment, especially from evaluators familiar with the psychosocial nuances of online communication.

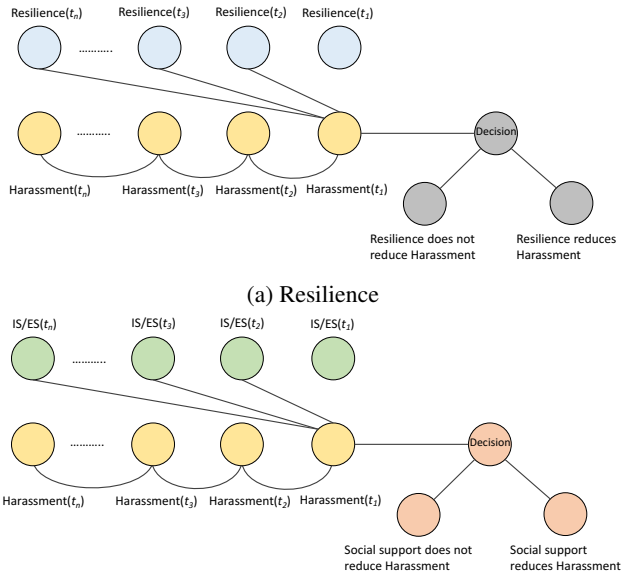
To this end, we employed a double-blind manual annotation study on a stratified random sample of 200 Instagram direct messages previously labeled by the classifier – balanced evenly between those marked as exhibiting resilience and those not. Two graduate researchers, each trained in qualitative content analysis and with prior exposure to literature on online resilience, independently annotated each message using a binary scheme: "resilient" versus "not resilient." To guide annotations, evaluators were provided with a rubric derived from the constructs of emotional strength, cognitive reframing, future orientation, and boundary-setting – dimensions consistently highlighted in the resilience literature (Layne et al. 2008; McLaughlin et al. 2021).

Inter-annotator reliability was high (Cohen's $\kappa = 0.74$), indicating substantial agreement and underscoring the clarity and applicability of the operationalization. Disagreements were resolved through discussion, facilitated by a senior co-author with clinical psychology training, allowing for consensus-based final labels. Comparison with the model's predictions yielded an overall accuracy of 0.89, precision of 0.91, recall of 0.87, and an F1-score of 0.89. These metrics suggest that the classifier is not only robust but also meaningfully aligned with human understanding of resilience as expressed in real-world conversational contexts.

Beyond quantitative validation, the annotation process also illuminated qualitative themes that the model appeared particularly adept at identifying. These included declarative refusals to tolerate disrespect, explicit statements of growth or insight post-trauma, and strategic emotional detachment from toxic interactions. However, subtler expressions—such as dry humor masking distress or implicit reappraisal remained more challenging for the classifier to detect reliably. These observations point to potential avenues for future model refinement, especially through the inclusion of multi-modal cues or context-aware modeling approaches.

Overall, this human evaluation showcases the classifier's value as a computational lens into youth expressions of resilience on social media, while highlighting the importance of continued reflexivity and iterative validation when deploying machine learning tools in socially and emotionally complex domains.

F. Granger Causality Framework



(a) Resilience

(b) Social (informational, emotional) support

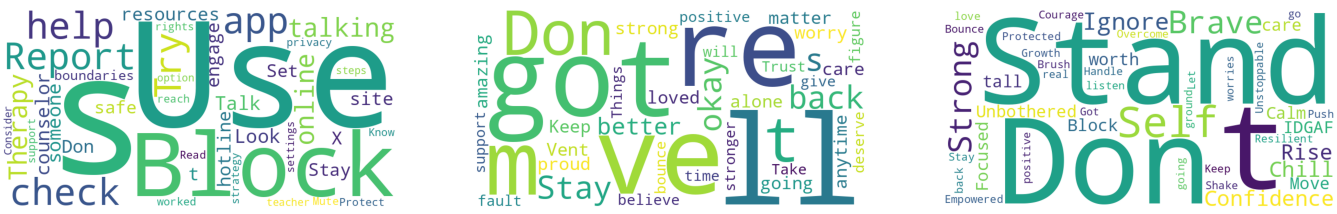
Figure A2: Granger causality based framework to assess the extent to which personal resilience and IS/ES received from others may help in reducing online harassment. It is noteworthy to mention that while Granger causality is a valuable tool for investigating causal relationships in time-series data, but it is important to remember that it establishes statistical, not necessarily causal, relationships. Additionally, it does not determine the direction of causality, only that there is a temporal relationship.

G: Qualitative Accounts of Support and Resilience in Response to Online Harassment

To enrich our quantitative findings around RQ3 and offer deeper insight into how mental health patients cope with online harassment, we conducted a thematic analysis of messages that scored high on resilience, emotional support, or informational support, as identified by the respective classifiers. Our objective was to qualitatively examine how resilience manifests in the language of patients navigating hostile Instagram interactions, and how social support may scaffold these expressions.

We sampled 100 such messages authored by mental health patients during periods of heightened harassment risk (e.g., post-hospitalization). Messages were reviewed independently by two trained researchers using an inductive coding strategy grounded in the resilience framework of McLaughlin et al. (2021). Three prominent themes emerged: *assertive boundary-setting*, *cognitive reframing*, and *strategic detachment*.

First, patients frequently articulated a shift in self-perception and agency through messages that set clear emotional or interpersonal boundaries. These often followed previous encounters with disrespect, suggesting a learning pro-



(a) Informational support (b) Emotional support (c) Resilience

Figure A3: Frequent terms in received and sent messages of patients, that scored high on IS and ES (a-b), and resilience (c).

cess rooted in trauma recovery. For example, one participant wrote (paraphrased), "*If she ever talks to me like that again, I'm walking away. I don't care who she is anymore.*" Such resistance not only indicates resilience but also demonstrates a recalibration of self-worth and tolerance for mistreatment.

Second, cognitive reframing appeared in reflective or metacognitive statements where patients reinterpreted their experience to reduce distress or enhance meaning-making. A (paraphrased) message read, "*I used to think I deserved this, but I'm realizing they lash out because of their own pain, not mine.*" Such statements reflect a psychological distancing that aligns with coping mechanisms, highlighting the role of inner dialogue in moderating emotional impact.

Third, strategic detachment manifested in dismissive or non-reactive tones, especially in conversations where patients deliberately avoided retaliatory engagement. (Paraphrased) messages like "*lol ok cool, think what you want*" or "*not gonna waste my energy on this bs*" reflect efforts to de-escalate even when provoked.

Notably, several peer messages expressed high emotional support. For instance, after a friend reassured a participant: "*you don't have to prove anything to them, you're doing great just as you are*" (paraphrased), the participant followed up with, "*thanks, I needed that. not letting this ruin my mood*" (paraphrased). This underscores the co-construction of resilience in interpersonal exchanges and the significance of validation in fostering recovery-oriented dialogue.

These qualitative findings suggest that resilience is not merely an individual trait but often emerges through relational dynamics and social feedback. In sum, this analysis shows that patients actively engage in meaning-making, assertiveness, and emotional boundary-setting as mechanisms for coping with harassment, and that these behaviors are often scaffolded by supportive messages from peers.

Appendix H: Comparative Analyses with Non-Hospitalized Mental Health Patients

To assess whether our findings uniquely characterize patients with a history of psychiatric hospitalization, we conducted additional analyses using a separate comparison group of individuals diagnosed with mental health conditions at an unrelated site – based on their medical record or ICD-10 code – but who were never hospitalized during the study period. This group (N = 67) was matched to the hospitalized group (N = 67) on age, gender distribution, and total message volume.

RQ1: Harassment Exposure. We compared average harassment levels across both groups. Per Table A2, hospitalized participants received a significantly higher proportion of harassment messages than their non-hospitalized peers. These differences underscore the heightened risk of interpersonal conflict faced by individuals following hospitalization.

Group	Mean Harassment (%)	SD	N
Hospitalized	12.5	3.2	67
Non-Hospitalized	8.3	2.7	67

Table A2: Mean harassment exposure by group.

RQ2: Temporal Harassment Dynamics. To evaluate whether harassment fluctuates meaningfully around key events, we replicated our temporal analysis for the non-hospitalized group using a pseudo-event (e.g., a spike in messaging volume or social engagement). As shown in Table A3, the hospitalized group showed a notable increase in harassment following hospitalization (+3.8%), whereas the non-hospitalized group exhibited minimal change (+0.3%). This suggests that hospitalization functions as a unique inflection point in terms of harassment exposure.

Group	Pre-Event (%)	Post-Event (%)	△ Change
Hospitalized	10.1	13.9	+3.8
Non-Hospitalized	8.2	8.5	+0.3

Table A3: Pre- and post-event harassment levels by group.

RQ3: Moderation by Support and Resilience. Finally, we assessed whether the buffering effects of social support and resilience observed in our main analyses were also evident among non-hospitalized mental health patients. Table A4 summarizes moderation results. Among hospitalized participants, emotional and informational support significantly reduced harassment risk, whereas these effects were weaker and non-significant in the non-hospitalized group. Resilience also showed a marginally protective effect for hospitalized individuals but not for others. These results suggest that protective mechanisms are more salient in contexts of acute vulnerability.

Summary. These comparative analyses confirm that the harassment risks and protective mechanisms highlighted in this study are specific to patients who have undergone

Moderator	Effect (Hosp.)	*p* (Hosp.)	Effect (Non-Hosp.)	*p* (Non-Hosp.)
ES	-0.27	0.004	-0.10	0.12
IS	-0.34	0.001	-0.12	0.07
Resilience	-0.12	0.090	-0.04	0.41

Table A4: Moderation effects of ES (emotional support), IS (informational support), and resilience on harassment.

psychiatric hospitalization. The absence of similar harassment spikes and weaker moderation effects among non-hospitalized individuals supports our interpretation that hospitalization marks a uniquely sensitive period requiring targeted online safety interventions.

Appendix I: Limitations of LLM-Generated Training Data and Inferred Scores

While our resilience classifier benefited from augmenting a small dictionary-based seed set with synthetic examples generated by an LLM (see Appendix D for more details), this approach carries limitations. First, LLM-generated data may introduce distributional biases, reflecting stylistic or cultural patterns of the model's training corpus rather than those of our target population. Second, synthetic examples may over-represent prototypical forms of resilience language while under-capturing the nuanced, context-specific ways youth express coping in private conversations. Third, the reliance on inferred scores – both for resilience and social support – means that these measures are approximations of latent psychological constructs, not direct ground-truth assessments. To mitigate these risks, we validated classifiers against a held-out set of human annotations (see Appendix E), but we emphasize that results should be interpreted cautiously. Future work should incorporate larger annotated corpora, participatory coding with lived-experience experts, and triangulation across multiple data modalities to strengthen construct validity.