Exploration of Digital Health Technologies

Open Exploration

Open Access Original Article

Analyzing tweets before and after Meta’s graphic self-harm imagery ban: a content analysis

Elnaz Moghimi[1,2,3]*, Kevin Keller[4], Sanjeef Thampinathan[2], William Cipolli[5], Hayden P. Smith[4]

[1]Waypoint Centre for Mental Health Care, Penetanguishene, ON L9M 1G3, Canada
[2]Faculty of Health Sciences, Queen’s University, Kingston, ON K7L 7X3, Canada
[3]Department of Psychiatry, University of Toronto, Toronto, ON M5T 1R8, Canada
[4]Department of Criminology and Criminal Justice, University of South Carolina, Columbia, SC 29208, USA
[5]Department of Mathematics, Colgate University, Hamilton, NY 13346, USA

***Correspondence:** Elnaz Moghimi, Waypoint Centre for Mental Health Care, Penetanguishene, ON L9M 1G3, Canada. elnaz.moghimi@queensu.ca

Academic Editor: Mirko Casu, University of Catania, Italy
Received: March 23, 2025 **Accepted:** May 27, 2025 **Published:** June 19, 2025

Cite this article: Moghimi E, Keller K, Thampinathan S, Cipolli W, Smith HP. Analyzing tweets before and after Meta’s graphic self-harm imagery ban: a content analysis. Explor Digit Health Technol. 2025;3:101154. https://doi.org/10.37349/edht.2025.101154

Abstract

Aim: The spread of suicide and non-suicidal self-injury (NSSI) content on social media has raised ongoing concerns about user safety and mental health. In response, social media platforms like Twitter (now X) and Meta (i.e., Facebook and Instagram) introduced content moderation policies to mitigate harm and promote safer digital environments. This study explored immediate trends in user discourse surrounding suicide and NSSI following the enactment of Meta’s graphic self-harm imagery ban. Specifically, it examined shifts in tweet tone, content type, and underlying themes immediately before and after the policy’s implementation.

Methods: A corpus of 3,846 tweets was analyzed. Within this corpus, tweets spanning 32 weeks from October 18, 2018, to May 29, 2019, were selected. These dates were chosen to encompass approximately 16 weeks before and after the enactment of the policy on February 7, 2019. Tweets were categorized according to slant, tweet category, and theme.

Results: The findings revealed notable shifts in online discourse. There was a significant decrease in the proportion of tweets identified as anti-self-harm tweets and a corresponding increase in the proportion of tweets aimed at understanding self-harm, many of which were coded as personal opinions or informative content. These trends suggest that while content promoting self-harm did not increase, the tone of discourse shifted toward greater nuance and reflection. This may reflect users’ growing efforts to process, contextualize, and share perspectives on self-harm in a policy-regulated environment.

Conclusions: Meta’s graphic self-harm imagery ban appeared to influence how users communicated about suicide and NSSI on Twitter, prompting more content centered on understanding and discussion. However, the findings also highlight challenges in balancing harm reduction with space for personal narratives. These insights emphasize the role of policy in shaping public discourse and the need for clear moderation strategies that distinguish harmful promotion from lived experience and peer support.

Keywords

Twitter, Meta, self-harm, suicide, social media content, content analysis, tweet

Introduction

According to the World Health Organization, more than 700,000 individuals die by suicide each year, making it the fourth leading cause of death amongst individuals aged 15 to 29 years [1]. Non-suicidal self-injury (NSSI) has also become a prevalent global issue, with reported prevalence rates ranging from 4.86% to 16.9% [2, 3]. NSSI is defined as inflicting intentional harm towards oneself but without ending one's life [4]. While there are many contributors associated with suicide and NSSI, social media is a predominant factor.

Recent studies have highlighted the complex relationship between social media use and mental health. Social media platforms can provide a sense of community and support for individuals struggling with mental health issues. However, excessive use and exposure to negative content can exacerbate feelings of loneliness, anxiety, and depression. The constant comparison to others' seemingly perfect lives can lead to low self-esteem and increased risk of NSSI and suicidal behaviors. Therefore, it is crucial to promote responsible social media use and provide resources for mental health support online [5].

Evidence suggests an association between social media use and self-injurious behaviors [6]. In a sample of 40,065 Norwegian university students from ages 18 to 35, those who actively used social media to post content had an increased risk of suicide and NSSI ideation [7]. Some reports attribute this risk to the contagion effect, where online social platforms facilitate the spread of thoughts, emotions, and behaviors across individuals and to the broader society [8]. Suicide and NSSI content can be spread and normalized via media exposure, ultimately leading to priming effects in vulnerable populations, including those without a history of these behaviors [9]. Accordingly, individuals active in online communities have reported learning about self-injury through these channels, and social learning has been linked to increases in suicide and NSSI [9, 10]. In a sample of individuals with recent episodes of NSSI, 43.6% reported learning about the behavior from peers or through media exposure [11].

Not surprisingly, social media platforms have been criticized for exposing users to content that promotes suicide and self-injury and a concomitant failure to provide safeguards that protect vulnerable populations. Platforms like Twitter (now known as "X") have offered a space for personal experiences and thoughts to be shared, including those pertaining to suicide and NSSI [12]. With more than 100 million users producing approximately 500 million tweets a day, content discussing suicide and NSSI is readily available [12]. A previous study categorized suicide and NSSI Twitter posts as manifestations of the behaviors, social responses to the behaviors, or exposure to the behaviors by mass communication [13]. Of these three responses, engagement in the behavior after exposure is the most concerning [14]. Meta-analytic data demonstrated that individuals who previously encountered content that promoted suicidal behaviour were more than three times more likely to die by suicide and almost three times as likely to survive a suicide attempt [15]. Therefore, it is important to consider the precariousness of populating social media platforms with suicide and NSSI content. Social media applications like Twitter have been the primary engagement zones for digital self-harm content, particularly for young people [16]. Twitter's shared interest groups allowed for anonymized communal conversations surrounding sensitive topics, creating an environment where users could be open without the fear of identification or judgment [17].

Currently, most major social media platforms have policies in place to regulate content related to suicide and NSSI. These platforms restrict explicit depictions of self-harm, or they may add warning labels to graphic imagery. A significant policy development occurred on February 7, 2019, when Meta (i.e., Instagram and Facebook) enacted a ban on graphic images that promoted self-harm [18]. The aim was to reduce the potential risk of normalizing or encouraging self-harm behaviors, while still allowing non-graphic content that could promote awareness, recovery, and connection to mental health resources. One year earlier, Twitter had implemented a similar policy, under which posts or messages that encouraged suicide or self-harm could be removed, and users could face account suspensions [19]. These collective efforts across platforms were designed to mitigate self-harm behaviors by limiting the promotion or glorification of harmful content.

The implementation of Meta's policy created a timely opportunity to examine how discourse around self-harm evolved in public online spaces. A prior study by Smith and Cipolli [20] used Twitter data to assess sentiment and topic shifts before and after the Instagram/Facebook ban, highlighting changes in emotional tone, such as increased anger and sadness and decreased joy and trust, following the policy. Their work emphasized the emotional resonance of the ban, revealing polarized responses that reflected both support for the effort and concerns about censorship or unintended consequences, however, it did not include a reference to the underlying themes driving these sentiments.

Building on this prior research, the current study extends the conversation beyond user sentiment by examining how content moderation policies enacted by one platform may shape norms across other social media environments through a process of "norming" that influences broader digital culture and governance [21]. Unlike Smith and Cipolli's focus on the emotional discourse surrounding the Instagram/Facebook ban, this study explores how such policy decisions reverberate across platforms, providing insight into how public discourse, across tone, topic, and underlying themes, may have been shaped immediately before and after its wake. The findings highlight how policy contributes to social media content and shapes public dialogue.

Materials and methods

Twitter data

The current study qualitatively analyzed a subset of tweets that were collected in a sentiment analysis and topic modelling study that programmatically explored content and sentiment in tweets mentioning a similar ban on graphic images of self-harm instituted by Facebook and Instagram [20]. Briefly, the premium Twitter API [22] captured 238,001 original posts, retweets, and quotes from June 1, 2018, to May 31, 2019, containing the terms "self-injurious behavior," "nonsuicidal self-injury," or "self-harm." These terms were selected based on previous studies on adolescent self-harm, and therefore may provide data more representative of academic- or policy-related discussion.

While previous work explored tweets mentioning "Facebook" or "Instagram" [20], the current study identified random samples of tweets that were posted in English, between October 18, 2018, to May 29, 2019. These dates were selected as they were approximately 16 weeks before and after Facebook and Instagram's graphic self-harm image ban was enacted on February 7, 2019 [18]. The rationale for selecting tweets within this range was to assess and compare the immediate impact of the policy on Twitter posts.

Sample size was calculated using a priori power analysis for a chi-squared test for comparing the rate of certain types of tweets before and after the ban was enacted. It was determined that a sample size of 99 tweets per week would yield adequate power (approximately 0.80) for detecting a small effect ($h = 0.1$). To account for the fact that many tweets do not have adequate information for coding (e.g., only links or emojis) and that some categories may be sparser than others, a random sample of 126 tweets per week (oversampled by roughly 25%) was taken for a total of 4,032 tweets. All initial data collection protocols were deemed exempt by the Colgate University Institutional Review Board (IRB) according to the Code of Federal Regulations Title 45, Part 46 (Exemption 2: Public data).

Data analysis

Tweets were aggregated and analyzed using an Excel file, with hashtags excluded to focus on the tweet content itself. The qualitative analysis was conducted using multi-stage content analysis methods [23, 24]. Two independent reviewers coded the tweets, and inter-coder reliability was assessed using Cohen's Kappa. Any discrepancies were resolved by a third reviewer. Tweets were initially classified by their slant—pro, anti, or neutral—to determine whether they conveyed a positive, negative, or neutral attitude toward self-harm or suicide.

Tweets were then classified into one of six content categories adapted from Jimenez-Sotomayor et al. [25].

(1) Personal opinions: express the author's subjective viewpoint, beliefs, or interpretations related to suicide or self-harm, often without citing evidence or sharing lived experience. These may include criticisms, support, or general thoughts on the topic.

(2) Informative: provide factual or educational content about self-harm or suicide, often referencing statistics, research findings, helpline resources, or mental health organizations.

(3) Jokes/ridicule: use humor, sarcasm, or mocking language to reference suicide or self-harm, often in ways that are dismissive, stigmatizing, or potentially harmful.

(4) Personal accounts/experiences: authors share their own lived experiences with self-harm or suicide, whether directly (e.g., describing their past or current struggles) or indirectly (e.g., reflecting on recovery or emotional states).

(5) Advice-seeking: author ask for help, guidance, or support in relation to their own or others' self-harm or suicidal thoughts. These often take the form of open-ended questions or direct appeals to the community.

(6) Other: do not clearly fall into any of the above categories. This may include ambiguous content, spam, song lyrics, memes, or general expressions unrelated to the main categories but still referencing suicide or self-harm in some way.

Lastly, manifest content analysis was used to identify themes—that is, patterns and meanings—across tweets [26]. This method provides a clear, observable, and literal description of the phenomenon by developing themes that remain close to the text [26, 27]. Initially, two independent coders analyzed approximately 2,500 tweets to create a codebook [28]. After discussion and revision among the research team, the codebook was applied to analyze the entire dataset. The final coding was reviewed by team members, and any discrepancies were resolved through discussion.

Descriptive statistics were used to report each tweet type's mean, median, standard deviation, and slant. To compare trends before and after the policy enactment, tweet slants and content categories from the period before the ban (October 18, 2018, to February 6, 2019) were compared to those posted after the ban (February 8, 2019, to May 29, 2019).

Results

Self-harm

For the analysis, suicide and NSSI behaviors were collectively reported as self-harm. From the 4,032 tweets, 186 were excluded (n = 143 not in English, n = 13 not relevant to self-harm, and n = 30 were on February 7th—the day the ban was announced). Subsequently, 3,846 tweets were analyzed. Among these tweets, the level of agreement when coding was very high for type (Cohen's k = 0.9294), slant (Cohen's k = 0.9004), and the combined coding (Cohen's k = 0.9018). Both before and after the enactment of the policy, the majority of tweets had an anti-self-harm slant, and only 2.78% of the tweets were pro-self-harm. Most of the tweets were personal opinions (28.19%), followed by personal accounts and experiences (25.20%) and informative posts (24.41%).

The most dominant themes included understanding self-harm (41.11%), political (16.56%), and support (13.36%). While anti-self-harm tweets were less prevalent after the ban, neutral self-harm tweets were more prevalent (Figure 1). Personal accounts and non-categorized tweets reduced after the ban (Figure 2). Conversely, tweets that were personal opinions, informative, and advice-seeking became more prevalent (Figure 2). The sole theme that became more prevalent after the ban was understanding self-harm (Figure 3). A full list of tweet slants, categories, and themes can be found in Table 1.

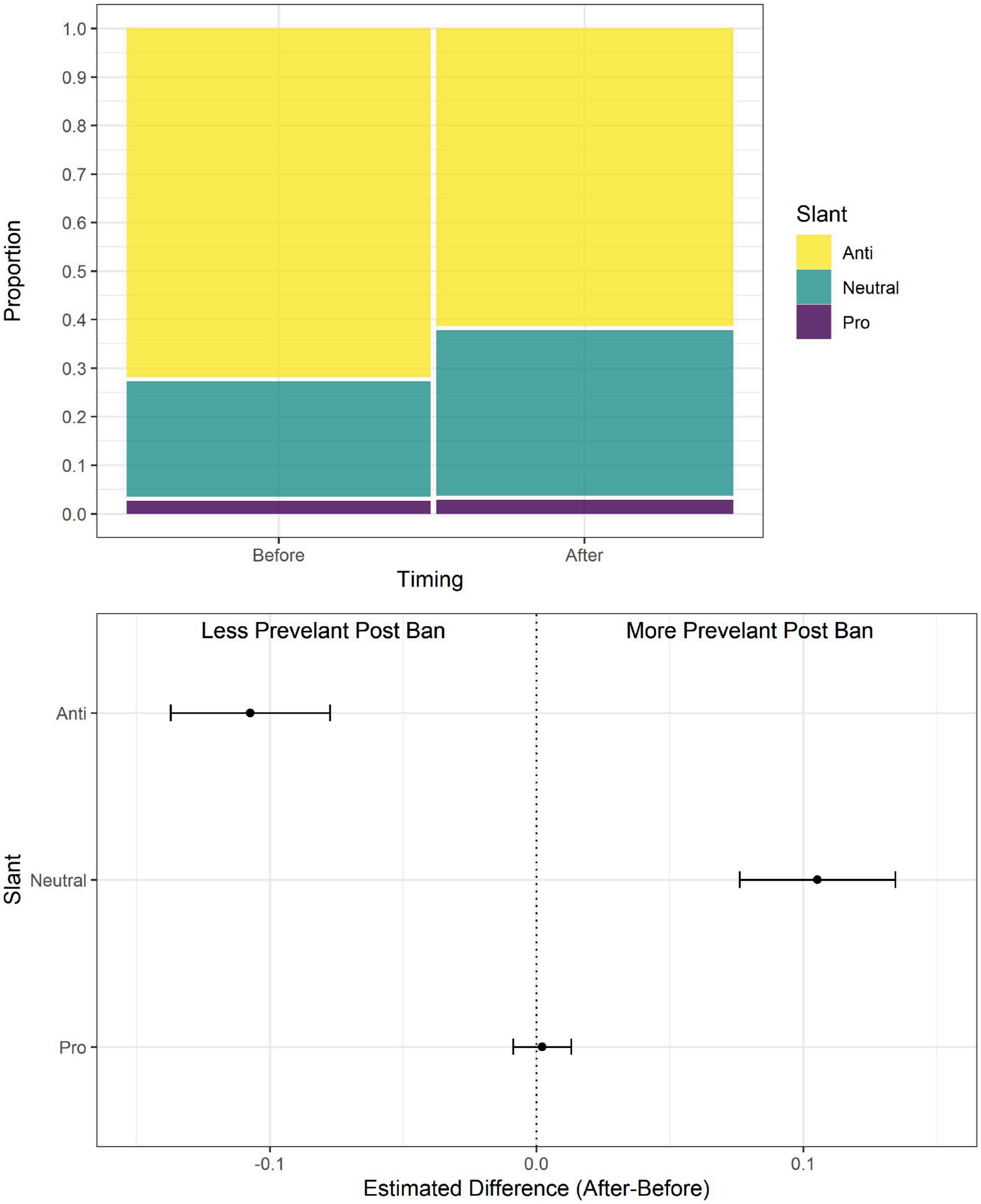

Figure 1. Tweet slant before and after the ban on graphic images of self-harm (top) and contrasts in prevalence (bottom)

Content analysis by theme

The qualitative analysis of the tweets revealed eleven distinct themes (Table 2). The majority of these tweets focused on understanding self-harm (41.11%), followed by political tweets that made inferences about self-harm (16.56%), supporting individuals affected by self-harm (13.36%), and calling out wrongdoings (12.69%). Only a small fraction of tweets promoted the acceptance of self-harm (0.16%).

Many tweets provided information to better understand what constitutes self-harm, who it may impact, and how it should be addressed. Most of these tweets were in discussion format (28%; 443/1,581), where different users shared information on topics related to self-harm. Many tweets also supported raising awareness (19.9%; 314/1,581) on self-harm. Information related to self-harm topics was also

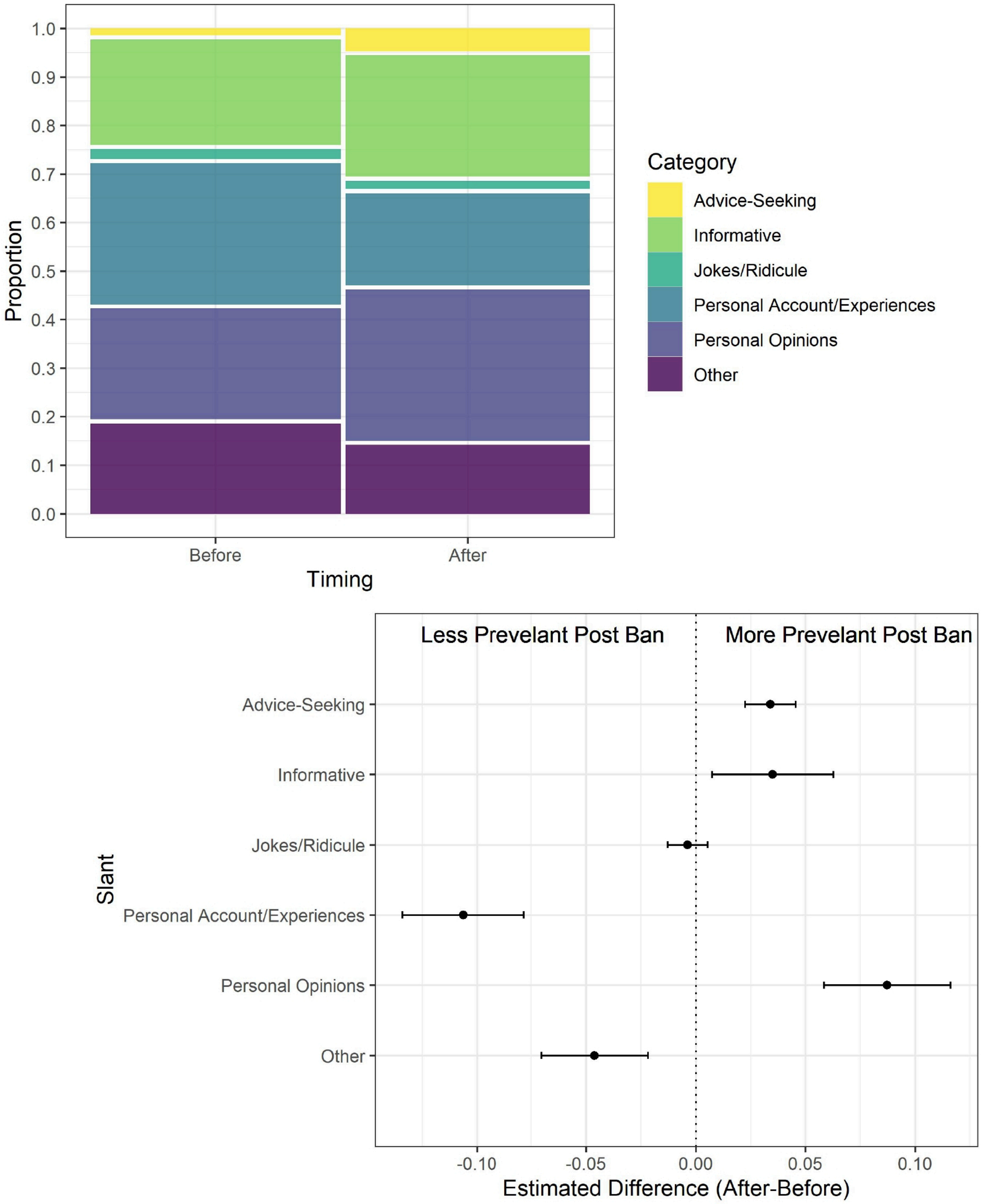

Figure 2. Tweet category before and after the ban on graphic images of self-harm (top) and contrasts in prevalence (bottom)

disseminated through links to news articles and blog posts (17.3%; 273/1,581). Personal experiences of users were sometimes used to learn about self-harm (15.5%; 245/1,581). Different reasons for engaging in self-harm were expressed by users. The most common reason was that the behavior was used as a coping mechanism (26.8%; 45/168). Urges or thoughts were also commonly expressed as a reason to self-harm (18.5%; 31/168). In some users, environmental factors, such as societal triggers and views on self-harm, also increased the risk of self-harm (8.9%; 15/168).

The most common comparisons of self-harm were different lifestyle choices (26.1%; 40/153), such as staying up too late or traveling to certain regions. Consuming certain types of foods, particularly fast foods, was also seen as self-harm (13.1%; 20/153). Viewing certain TV shows and video games was also described

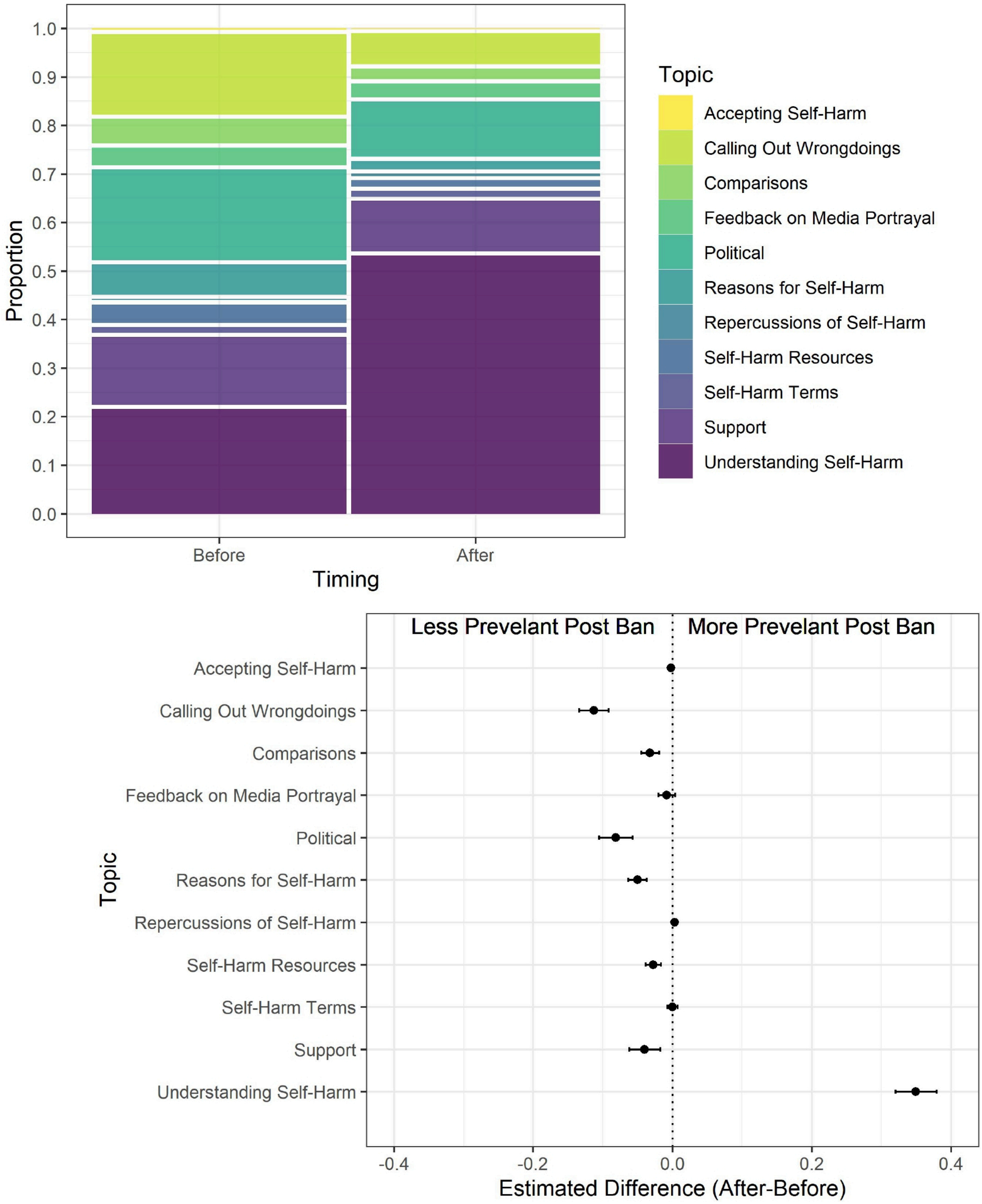

Figure 3. Tweet theme before and after the ban on graphic images of self-harm (top) and contrasts in prevalence (bottom)

as a form of self-harm (10.5%; 16/153). Much of the commentary about the media centered on self-harm portrayed in TV or film (39.1%; 52/133), social media (30.1%; 40/133), and books (14.3%; 19/133). Many of the tweets supported or challenged censorship of self-harm and highlighted the importance of discussing self-harm in these media formats.

Many users also expressed what they perceived to be wrongdoings. The majority of these tweets focused on social media (15%; 73/488), with some users concerned about the lack of attention that social media companies pay to self-harm content and users reporting this content. Some users (13.9%; 68/488) also discussed the quality of care that made it difficult for individuals to seek help for self-harm. Other concerns that were expressed pertained to reducing bullying that may result in self-harm (12.7%; 62/488),

Table 1. Tweet counts, organized by timing, slant, and type. The chi-squared test evaluates whether the proportion of tweets differs pre-ban and post-ban for each slant, category, and theme

Information	Total *n* (%)	Pre-ban	Post-ban	χ^2 (df), adjusted *p*
Slant				
Anti	2,611 (67.89)	1,423	1,188	50.37 (1), < 0.001
Neutral	1,128 (29.33)	469	659	50.85 (1), < 0.001
Pro	107 (2.78)	52	55	0.10 (1), 0.796
Category				
Advice-seeking	120 (3.12)	28	92	35.58 (1), < 0.001
Informative	939 (24.41)	441	498	6.19 (1), 0.018
Jokes/ridicule	72 (1.87)	40	32	0.55 (1), 0.511
Personal account/experiences	969 (25.20)	592	377	57.09 (1), < 0.001
Personal opinions	1,084 (28.19)	464	620	35.76 (1), < 0.001
Other	662 (17.21)	379	283	14.06 (1), < 0.001
Theme				
Accepting self-harm	6 (0.16)	5	1	1.44 (1), 0.271
Calling out wrongdoings	488 (12.69)	355	133	109.18 (1), < 0.001
Comparisons	153 (3.98)	108	45	24.78 (1), < 0.001
Feedback on media portrayal	133 (3.46)	75	58	1.65 (1), 0.249
Political	637 (16.56)	400	237	45.23 (1), < 0.001
Reasons for self-harm	168 (4.37)	133	35	56.38 (1), < 0.001
Repercussions of self-harm	12 (0.31)	3	9	2.20 (1), 0.184
Self-harm resources	106 (2.76)	80	26	26.08 (1), < 0.001
Self-harm terms	48 (1.25)	24	24	0 (1), 1
Support	514 (13.36)	298	216	12.76 (1), < 0.001
Understanding self-harm	1,581 (41.11)	463	1,118	484.0 (1), < 0.001

Table 2. Content analysis of tweets resulted in the emergence of 11 themes, organized in ascending order of frequency

Theme	Description	Total # of tweets, *n* (%)
Understanding self-harm	Tools, research, and/or resources to better understand self-harm	1,581 (41.11)
Political	Self-harm terms used in political information, commentary, or discussions	637 (16.56)
Support	Individual support given to or received by those affected by self-harm	514 (13.36)
Calling out wrongdoings	Discouraging or pointing out the negative elements of self-harm and/or its promotion	488 (12.69)
Reasons for self-harm	Specific factors that encourage engagement in self-harm	168 (4.37)
Comparisons	Self-harm was likened to lifestyle, objects, food, and certain behaviors	153 (3.98)
Feedback on media portrayal	Reflections on portrayals of self-harm in print and digital media	133 (3.46)
Self-harm help resource	Community services and resources for those affected by self-harm	106 (2.76)
Minor themes		
Self-harm terms	Single-word posts that pertain to self-harm, without a coherent sentence structure	48 (1.25)
Repercussions of self-harm	Description(s) of the negative impacts of self-harm behaviors	12 (0.31)
Accepting self-harm	Explicit promotion and/or glorification of self-harm behaviors	6 (0.16)

trivializing or minimizing self-harm behaviors (9.8%; 48/488), and glorifying self-harm (7.8%; 38/488). Most political tweets considered certain policies and laws as being a form of self-harm. More than half of the political tweets centered on either the United Kingdom (18.1%; 115/637) or, more specifically, Brexit (51.2%; 326/637), referring to it as an act of self-harm.

Tweets pertaining to support involved users being proud of themselves (39.7%; 204/514) and others (31.7%; 163/514) for coping with self-harm and their recovery journey. Several encouraging tweets mentioned the duration of time that individuals refrained from engaging in self-harm. Some users also expressed the need to get help from others (14%; 72/514). Links to resources for individuals seeking help from self-harm were predominantly treatments or services (34.9%; 37/106) or hotlines (35.8%; 38/106) provided by organizations focused on self-harm. Links to mobile applications to manage self-harm or related mental health concerns, such as anxiety, panic attacks, and depression, were also observed in the corpus of tweets (11.3%; 12/106).

With respect to the minor themes, the majority of tweets accepting self-harm promoted the behavior and glorified self-harm as a positive activity that should be engaged in. Tweets pertaining to self-harm terms contained either a single word or multiple words related to self-harm, and without a coherent sentence structure.

Tweet slants and categories within themes

With respect to tweet slants (Figure 4), most of the anti-self-harm tweets (40.4%; 1,055/2,611) fell under the theme of understanding self-harm. The same theme also had a large proportion of the total neutral slanted self-harm tweets (43.4%; 490/1,128). While there were very few pro-self-harm tweets, most fell under the theme of understanding self-harm (33.6%; 36/107) and reasons for self-harm (29.9%; 32/107).

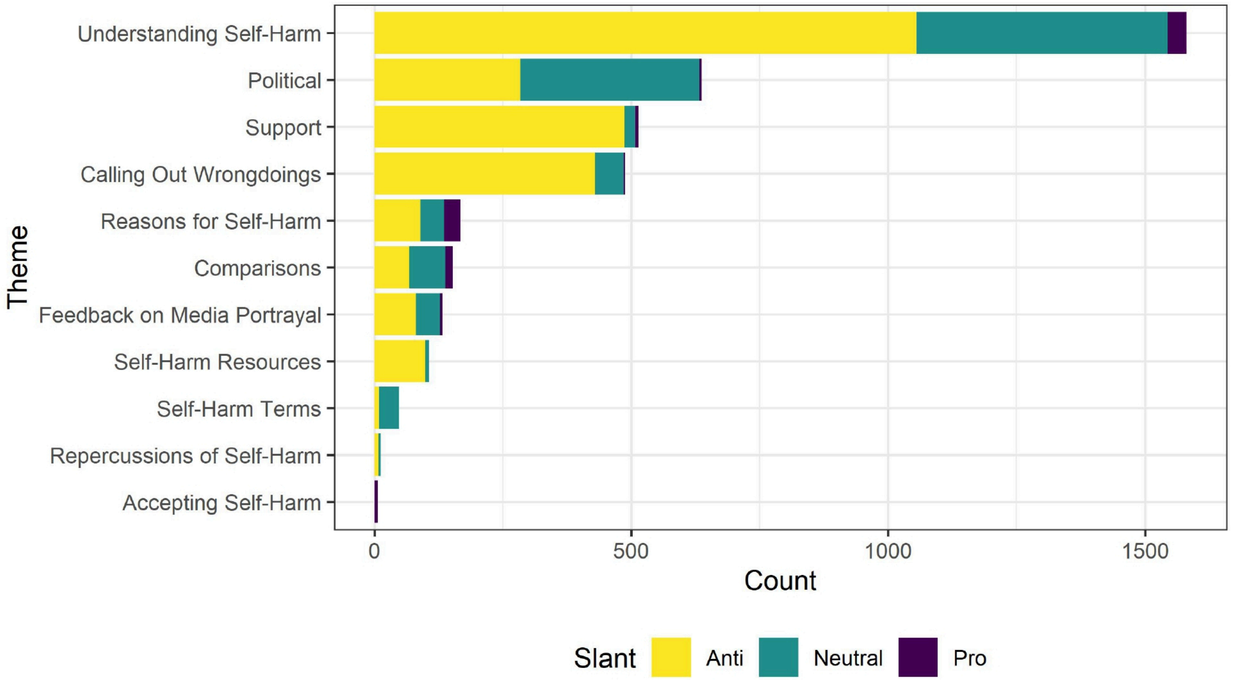

Figure 4. Tweet slant, organized by theme

Within almost all the themes, anti-self-harm tweets were the most predominant. Tweets categorized under political (44.6%; 284/637), comparisons (44.4%; 68/153), and self-harm terms (81.3%; 39/48) were mostly neutral slanted. All themes under accepting self-harm were pro-self-harm tweets (*n* = 6). Conversely, none of the tweets that fell under the themes of self-harm help resource (*n* = 106), self-harm terms (*n* = 48), and repercussions of self-harm (*n* = 12) contained pro-self-harm tweets.

Regarding tweet categories (Figure 5), most personal opinions fell under the theme of understanding self-harm (38.5%; 417/1,084) or calling out wrongdoings (25.4%; 275/1,084). More than half of informative tweets were also categorized as understanding self-harm (77.6%; 729/939). Half of the tweets that were jokes or ridicule were comparisons (50%; 36/72). Personal accounts and experiences were mostly coded under the themes of support (35.2%; 341/969) or understanding self-harm (32.5%; 315/969). Tweets under advice-seeking mostly represented the theme of understanding self-harm (46.7%; 56/120) or support (29.2%; 35/120). Tweets that did not fall under any type (other) were mostly political (79.6%; 527/662).

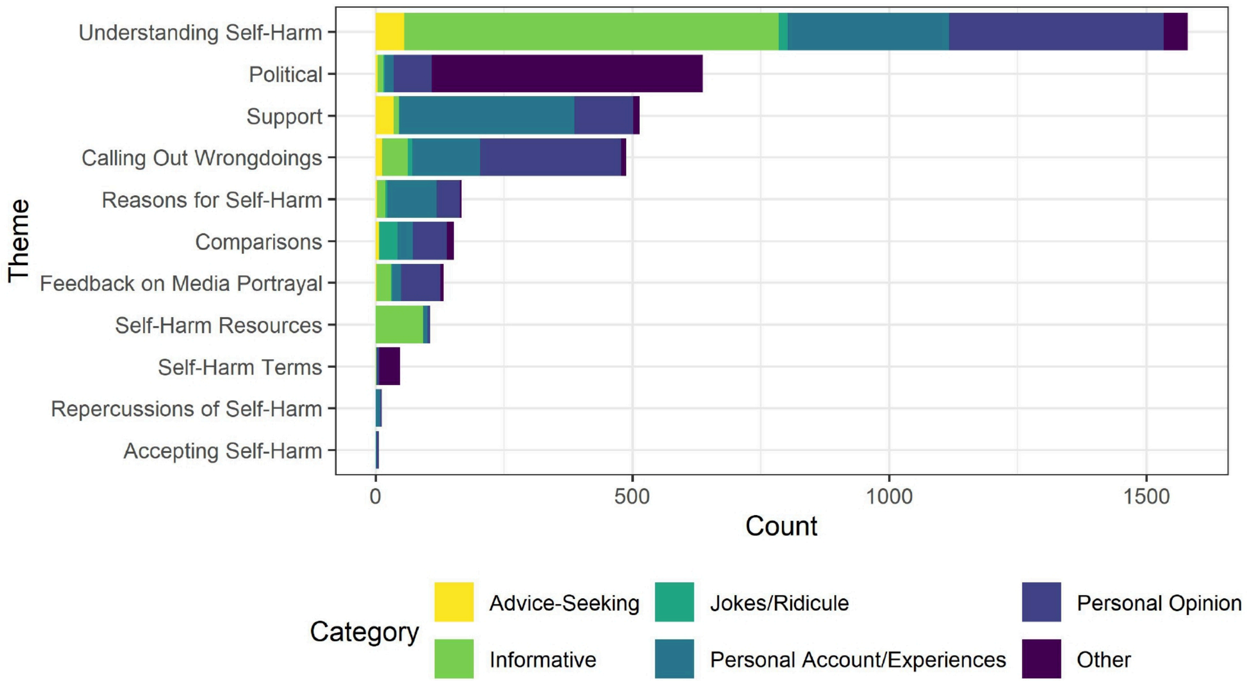

Figure 5. Tweet category, organized by theme

Discussion

The current study aimed to develop an understanding of the types of tweets surrounding Meta's graphic self-harm imagery ban. The findings indicated that the majority of tweets promoted a better understanding of self-harm, had an anti-self-harm slant, and were personal opinions. Personal opinions and informative tweets comprised the largest proportion of tweets intended to enhance understanding of self-harm. This finding highlights the importance of considering the types of information disseminated on the social media platform.

At the same time, the observed decrease in anti-self-harm tweets following the Instagram and Facebook ban may reflect a broader shift in user behavior and emerging norms across platforms, including Twitter. This change could signal a growing reluctance among users to engage explicitly with self-harm content, particularly content that critiques or discourages the behavior, due to concerns about visibility, censorship, or being algorithmically flagged. While messages discouraging self-harm remained prevalent, their reduced volume suggests that users may be adapting how they communicate about sensitive topics in response to evolving moderation policies. In parallel, there was a notable increase in posts framed as personal reflections or informational content, indicating a discursive shift toward more indirect, explanatory, or narrative-based engagement. These trends suggest that users are actively recalibrating the language and tone of self-harm discourse in ways that may reflect an emerging norming effect stemming from platform-level policy changes.

Notably, tweets that explicitly accepted or promoted self-harm were extremely rare in both time periods and appeared to decline further following the implementation of the ban. Although this reduction was not statistically significant, likely due to the small number of tweets in this category, the downward trend aligns with the policy's intended objectives and may reflect improvements in content moderation or increased user self-censorship. Despite their rarity, tweets that endorse or normalize self-harm raise important ethical concerns. Even infrequent exposure to such messages can contribute to harm contagion, particularly for vulnerable users who may interpret them as validating or encouraging self-injury [8]. As such, social media platforms must remain vigilant not only in identifying this content but also in ensuring it is contextualized and addressed in ways that prioritize user safety and mental health.

Considering that a large proportion of social media users obtain their information online, the predominance of personal opinions may result in seemingly informative tweets being misleading or inaccurate. Indeed, the World Health Organization has recently coined the term infodemic, characterized by an overwhelming influx of information, including false or misleading content in both digital and physical realms, leading to confusion and potentially harmful health-related behaviors [29]. Although infodemics have been defined in the context of disease outbreaks, their relevance to self-harm has been demonstrated in the current study.

While the number of neutral tweets significantly increased after policy enactment, there was also a simultaneous increase in both informative tweets and personal opinions. Although the policy may have encouraged discourse surrounding a better understanding of self-harm, the information also has the potential to promote a skewed understanding of the topic. Future studies should consider how different types of tweets aimed at improving a user's understanding of self-harm are perceived by users, particularly those vulnerable to suicide and NSSI.

Despite concerns about how information is disseminated on Twitter, the high level of support expressed on the platform may provide insight into why many users may feel a sense of community and peer support [30, 31]. The current study shed light on the dynamics of support—namely, that it provides a space for individuals to connect and support one another by sharing personal experiences with self-harm, providing resources to seek care, expressing specific reasons why self-harm may occur, and offering feedback on how these behaviors are portrayed in the media. While nearly all users called out the negative implications associated with self-harm, less than 1% posted pro-self-harm tweets that fell into this category. Many of these tweets accused social media of censoring their thoughts and opinions related to self-harm.

In line with previous findings from the study authors and others, political tweets disseminated on the platform focused heavily on conflating Brexit with self-harm [20, 32]. Nearly all of these tweets fell under none or more than one type and were either anti- or neutral-slanted. While the impact of calling social issues acts of self-harm has not been assessed, tweets that draw comparisons to self-harm may perpetuate mental health stigma [33]. It is critical for future work to determine if reporting Brexit in this manner may have contributed, at least to some degree, to the raised suicide risk in Brexit-voting communities [34] and increased anxiety and compromised mental health in migrants residing in the United Kingdom [35]. It certainly raises the intriguing linguistic finding of this study, where the term self-harm extended beyond traditional definitions of the behavior.

Limitations

A notable limitation of the study pertained to the self-harm language that was captured. Although the dissemination of self-harm content is a pervasive problem on social media, only about 3% of the total tweets analyzed were pro-self-harm. This number is substantially lower than reports of 9–66% pro-self-harm content within other social media platforms [36] but may also reflect the impacts of the earlier ban on Twitter. In contrast, a recent report indicated a 500% increase in self-harm-related hashtags since October 2022 [16]. Specifically, they noted that the "#shtwt" (self-harm TWiTter) hashtag, grew from 4k to nearly 39k tweets in less than a year despite moderation policies. Further, they describe the development of ban-evading coded language (e.g., "cat scratches" instead of "cuts") that enables users to sidestep moderation. See Atauri-Mezquida et al. [37] for more discussion about these workarounds and further investigations into the (lack of) effectiveness of the ban and Twitter's moderation policy implementations

The low number of pro-self-harm tweets in the current study may also indicate that the ambiguous language of self-harm within social media platforms evolves in ways to bypass self-harm policies [38] (e.g., #shtwt and "cat scratches"). Given the dynamic and ever-changing nature of this language, it may be challenging to delineate unique terms and examples over the long term. Although identifying specific strategies for tracking emergent self-harm language was beyond the scope of this study, future research should consider integrating machine learning, user-informed keyword development, or collaboration with people with lived experience to better detect and interpret evolving terminology.

Conclusion

The current study demonstrated observable changes in online dialogue surrounding self-harm, offering insight into how public discourse may be shaped by platform-level policies aimed at reducing exposure to potentially harmful content. While the implementation of Instagram and Facebook's ban on graphic self-harm imagery appeared to coincide with a shift toward conversations focused on understanding self-harm, it also highlighted the challenge of distinguishing between factual information and personal narratives. Importantly, personal reflections and firsthand accounts may serve a meaningful peer-support function, particularly for users seeking connection, validation, or a sense of community. Although this study was not designed to evaluate regulatory or governmental interventions, the findings may inform platform governance strategies and contribute to broader public health conversations. Future research may benefit from interdisciplinary partnerships with mental health organizations, legal scholars, and user communities to guide the development of user-centered, evidence-informed content moderation practices.

Abbreviations

NSSI: non-suicidal self-injury

Declarations

Author contributions

EM: Conceptualization, Formal analysis, Writing—original draft, Writing—review & editing. KK and ST: Formal analysis, Writing—original draft, Writing—review & editing. WC: Conceptualization, Formal analysis, Writing—review & editing. HPS: Conceptualization, Writing—review & editing.

Conflicts of interest

The authors declare that they have no conflicts of interest.

Ethical approval

All initial data protocols were deemed exempt by the Colgate University Institutional Review Board (IRB) according to the Code of Federal Regulations Title 45, Part 46 (Exemption 2: Public data).

Consent to participate

Not applicable.

Consent to publication

Not applicable.

Availability of data and materials

The datasets and code analyzed for this study will be made available via [https://github.com/WilliamCipolli/EDHT-Moghimi-et-al-2025.git].

Funding

Not applicable.

Copyright

Publisher's note

Open Exploration maintains a neutral stance on jurisdictional claims in published institutional affiliations and maps. All opinions expressed in this article are the personal views of the author(s) and do not represent the stance of the editorial team or the publisher.

References

1. Suicide [Internet]. World Health Organization; c2025 [cited 2023 Nov 9]. Available from: https://www.who.int/news-room/fact-sheets/detail/suicide
2. Gillies D, Christou MA, Dixon AC, Featherston OJ, Rapti I, Garcia-Anguita A, et al. Prevalence and Characteristics of Self-Harm in Adolescents: Meta-Analyses of Community-Based Studies 1990–2015. Journal of the American Academy of Child & Adolescent Psychiatry. 2018;57:733–41. [DOI]
3. Liu RT. The epidemiology of non-suicidal self-injury: lifetime prevalence, sociodemographic and clinical correlates, and treatment use in a nationally representative sample of adults in England. Psychol Med. 2023;53:274–82. [DOI] [PubMed] [PMC]
4. Conner BT, Kentopp SD, O'Donnell MB, Wallace GT, Morse JL, Arkfeld PA, et al. Meaning in Life Moderates Relations between Personality and Temperament and Nonsuicidal Self-Injury in Hospitalized Adolescents. J Youth Adolesc. 2022;51:1622–35. [DOI] [PubMed]
5. Smith KE, Acevedo-Duran R, Lovell JL, Castillo AV, Cardenas Pacheco V. Youth Are the Experts! Youth Participatory Action Research to Address the Adolescent Mental Health Crisis. Healthcare (Basel). 2024;12:592. [DOI] [PubMed] [PMC]
6. Nesi J, Burke TA, Bettis AH, Kudinova AY, Thompson EC, MacPherson HA, et al. Social media use and self-injurious thoughts and behaviors: A systematic review and meta-analysis. Clin Psychol Rev. 2021;87:102038. [DOI] [PubMed] [PMC]
7. Kingsbury M, Reme BA, Skogen JC, Sivertsen B, Øverland S, Cantor N, et al. Differential associations between types of social media use and university students' non-suicidal self-injury and suicidal behavior. Computers in Human Behavior. 2021;115:106614. [DOI]
8. Seong E, Noh G, Lee KH, Lee JS, Kim S, Seo DG, et al. Relationship of Social and Behavioral Characteristics to Suicidality in Community Adolescents With Self-Harm: Considering Contagion and Connection on Social Media. Front Psychol. 2021;12:691438. [DOI] [PubMed] [PMC]
9. Jarvi S, Jackson B, Swenson L, Crawford H. The Impact of Social Contagion on Non-Suicidal Self-Injury: A Review of the Literature. Arch Suicide Res. 2013;17:1–19. [DOI] [PubMed]
10. Syed S, Kingsbury M, Bennett K, Manion I, Colman I. Adolescents' knowledge of a peer's non-suicidal self-injury and own non-suicidal self-injury and suicidality. Acta Psychiatr Scand. 2020;142:366–73. [DOI] [PubMed]
11. Heath NL, Toste JR, Nedecheva T, Charlebois A. An Examination of Nonsuicidal Self-Injury Among College Students. Journal of Mental Health Counseling. 2008;30:137–56. [DOI]
12. Alhassan MA, Pennington D. Investigating non-suicidal self-injury discussions on Twitter. International Conference on Social Media and Data Mining-ICSMDM; 2021.
13. Silva AC, Vedana KGG, dos Santos JCP, Pillon SC, Ventura CAA, Miasso AI. Analysis of nonsuicidal self-injury posts on Twitter: A quantitative and qualitative research. Research, Society and Development. 2021;10:e40410413017. [DOI]
14. Muehlenkamp J, Brausch A, Quigley K, Whitlock J. Interpersonal Features and Functions of Nonsuicidal Self-injury. Suicide Life Threat Behav. 2013;43:67–80. [DOI] [PubMed]
15. Hill NTM, Robinson J, Pirkis J, Andriessen K, Krysinska K, Payne A, et al. Association of suicidal behavior with exposure to suicide and suicide attempt: A systematic review and multilevel meta-analysis. PLoS Med. 2020;17:e1003074. [DOI] [PubMed] [PMC]
16. Online Communities of Adolescents And Young Adults Celebrating, Glorifying, and Encouraging Self-Harm and Suicide are Growing Rapidly on Twitter [Internet]. Network Contagion Research Institute; c2025 [cited 2023 Nov 17]. Available from: https://networkcontagion.us/reports/8-29-22-online-communities-of-adolescents-and-young-adults-celebrating-glorifying-and-encouraging-self-harm-and-suicide-are-growing-rapidly-on-twitter/
17. Juwita ET, Effendi AZ, Pandin MGR. The Effect of Anonymity on Twitter towards its Users Based on Derek Parfit's Personal Identity Theory. 2021. [DOI]

18. Changes We’re Making to Do More to Support and Protect the Most Vulnerable People who Use Instagram [Internet]. Instagram Blog; 2019 [cited 2025 May 1]. Available from: https://about.instagram.com/blog/announcements/supporting-and-protecting-vulnerable-people-on-instagram
19. Suicide and Self-harm policy [Internet]. X Corp.; c2025 [cited 2023 Apr 11]. Available from: https://help.twitter.com/en/rules-and-policies/glorifying-self-harm
20. Smith H, Cipolli W. The Instagram/Facebook ban on graphic self-harm imagery: A sentiment analysis and topic modeling approach. Policy & Internet. 2022;14:170–85. [DOI]
21. Dempsey RC, Fedorowicz SE, Wood AM. The role of perceived social norms in non-suicidal self-injury and suicidality: A systematic scoping review. PLoS One. 2023;18:e0286118. [DOI] [PubMed] [PMC]
22. Python client for the Twitter search endpoints. GitHub, Inc.; c2025 [cited 2025 May 1]. Available from: https://github.com/twitterdev/search-tweets-python
23. Krippendorff K. Content analysis: An introduction to its methodology. Sage Publications; 2018.
24. Neuendorf KA. Content analysis and thematic analysis. In: Advanced Research Methods for Applied Psychology. Routledge; 2018.
25. Jimenez-Sotomayor MR, Gomez-Moreno C, Soto-Perez-de-Celis E. Coronavirus, Ageism, and Twitter: An Evaluation of Tweets about Older Adults and COVID-19. J Am Geriatr Soc. 2020;68:1661–5. [DOI] [PubMed] [PMC]
26. Kleinheksel AJ, Rockich-Winston N, Tawfik H, Wyatt TR. Demystifying Content Analysis. Am J Pharm Educ. 2020;84:7113. [DOI] [PubMed] [PMC]
27. Bengtsson M. How to plan and perform a qualitative study using content analysis. NursingPlus Open. 2016;2:8–14. [DOI]
28. Hsieh HF, Shannon SE. Three Approaches to Qualitative Content Analysis. Qual Health Res. 2005;15: 1277–88. [DOI] [PubMed]
29. Infodemic [Internet]. World Health Organization; c2025 [cited 2024 Jan 10]. Available from: https://www.who.int/health-topics/infodemic/understanding-the-infodemic-and-misinformation-in-the-fight-against-covid-19#tab=tab_1
30. Hilton CE. Unveiling self-harm behaviour: what can social media site Twitter tell us about self-harm? A qualitative exploration. Journal of Clinical Nursing. 2017;26:1690–704. [DOI] [PubMed]
31. Thorn P, La Sala L, Hetrick S, Rice S, Lamblin M, Robinson J. Motivations and perceived harms and benefits of online communication about self-harm: An interview study with young people. Digital Health. 2023;9:20552076231176689. [DOI] [PubMed] [PMC]
32. Shanahan N, Brennan C, House A. Self-harm and social media: thematic analysis of images posted on three social media sites. BMJ Open. 2019;9:e027006. [DOI] [PubMed] [PMC]
33. Corrigan PW, Powell KJ, Michaels PJ. The Effects of News Stories on the Stigma of Mental Illness. J Nerv Ment Dis. 2013;201:179–82. [DOI] [PubMed]
34. Steeg S, Webb RT, Ibrahim S, Appleby L, Kapur N. Suicide rates and voting choice in the UK’s 2016 national Brexit referendum on European Union membership: cross-sectional ecological investigation across England’s local authority populations. BJPsych Open. 2020;6:e57. [DOI] [PubMed] [PMC]
35. Frost DM. Hostile and harmful: Structural stigma and minority stress explain increased anxiety among migrants living in the United Kingdom after the Brexit referendum. J Consult Clin Psychol. 2020;88: 75–81. [DOI] [PubMed]
36. Picardo J, McKenzie SK, Collings S, Jenkin G. Suicide and self-harm content on Instagram: A systematic scoping review. PLoS One. 2020;15:e0238603. [DOI] [PubMed] [PMC]
37. Atauri-Mezquida D, Nogales-González C, Martínez-Pastor E. Exploring self-harm on Twitter (X): Content moderation and its psychological effects on adolescents. Online Journal of Communication and Media Technologies. 2025;15:e202503. [DOI]
38. Moreno MA, Ton A, Selkie E, Evans Y. Secret Society 123: Understanding the Language of Self-Harm on Instagram. J Adolesc Health. 2016;58:78–84. [DOI] [PubMed] [PMC]