Online Journal of Communication and Media Technologies, 2025, **15**(1), e202503
e-ISSN: 1986-3497

OPEN ACCESS

Research Article

Exploring self-harm on Twitter (X): Content moderation and its psychological effects on adolescents

David Atauri-Mezquida [1]
0000-0001-8451-5746

Celia Nogales-González [2]
0000-0002-5269-6930

Esther Martínez-Pastor [3*]
0000-0002-2861-750X

[1] University Rey Juan Carlos, Móstoles, SPAIN
[2] University Rey Juan Carlos, Alcorcón, SPAIN
[3] University Rey Juan Carlos, Fuenlabrada, SPAIN
[*] Corresponding author: esther.martinez.pastor@urjc.es

Citation: Atauri-Mezquida, D., Nogales-González, C., & Martínez-Pastor, E. (2025). Exploring self-harm on Twitter (X): Content moderation and its psychological effects on adolescents. *Online Journal of Communication and Media Technologies, 15*(1), e202503. https://doi.org/10.30935/ojcmt/15867

ARTICLE INFO

Received: 7 Aug 2024

Accepted: 7 Nov 2024

ABSTRACT

The number of non-suicidal self-injuries (NSSI) has grown exponentially in the last two decades, especially in the young population. Likewise, the number of textual tweets and images in photography and video has increased, to share the experience of self-harm. A quantitative and qualitative analysis of the posts on self-harm on Twitter is presented. The objective is to make visible the importance of this growing phenomenon among young people and to discuss its psychological and behavioral impact on the individual, as well as to analyze whether Twitter has adequate content controls in accordance with its conduct policies. 32,231 tweets were collected between 24 November 2022, and 29 January 2023, containing four keywords: "selfharm", "shtwt", "goretwt", and "ouchietwt". An average of 725 daily tweets were found. These tweets were posted by a total of 11,749 different users. Of those published, only 3,859 tweets (8.3%) were blocked by Twitter's content moderation procedures, and they did so only after having produced a high number of impressions and reactions in the community. No coherence was found between the tweets blocked due to their sensitive or non-sensitive load, or between Twitter's criteria for blocking tweets and those that are finally deleted. By number of posts, the violence of the messages, photographs, videos, and the interactions produced because of these, the importance of the procedures of moderation and supervision of content by Twitter are discussed, in terms of the danger and the psychological impact on the users.

Keywords: self-harm (NSSI), X, Twitter, content moderation, psychological impact, shtwt

INTRODUCTION

The Phenomenon of Non-Suicidal Self-Injury

Non-suicidal self-injury (NSSI) has been defined as a person's self-inflicted damage to body tissues, caused consciously and deliberately, to modify their mood and without suicidal intent (Pérez-Elizondo, 2020). Furthermore, they have a self-soothing nature (Nock, 2010; Nock & Favazza, 2009), they are repetitive, direct and with a low degree of lethality (Favazza & Favazza, 1987; Pattison & Kahan, 1983). This behavior is not new; in past decades, NSSI was linked to mental health issues, sexual abuse, and domestic violence (Faura-García,

Note: The fieldwork for this study was conducted when the social network was called Twitter, but since July 2023, it has been renamed X. This study keeps the name Twitter along the text.

Calvete & Orue, 2021; Sutton, 2007). In 2014, the diagnostic and statistical manual of mental disorders (DSM-5) (American Psychiatric Association, 2014) included NSSI as a behavior with its own entity, warranting further study and monitoring by health professionals. Currently, it is considered a mental disorder and should "be a focus of clinical attention" (DSM-5 TR) (American Psychiatric Association, 2022).

Furthermore, NSSI is a behavior disapproved of by society and highly stigmatized (Staniland et al., 2021). NSSI produces habituation to pain, which encourages self-injury to become increasingly intense, in order to be able to feel the same as what was previously felt with less intense self-injury (Stacy et al., 2018). In recent years, NSSI have increased among minors and young people aged 15 to 19, as indicated by UNICEF (2020, 2021a, 2021b). This increase, especially in adolescence, indicates a public health problem that needs to be studied and solved (Barrocas et al., 2012).

NSSI are also considered an addictive behavior (Pérez-Elizondo, 2021). However, shortly, the stressors that led to such an act reappear and lead the person to a new self-harm episode. NSSI is a maladaptive and avoidance-oriented coping strategy for stress, it seek short-term escape of discomfort (Angelakis & Gooding, 2021). Self-injury includes many diverse topographies, such as cutting, burning, scratching strongly, pulling out hair, breaking bones, hitting oneself, intentionally aggravating a wound, etc. The most typical way is to cut the skin in different areas, especially on the arms and thighs (Klonsky, 2011; Klonsky & Olino, 2008).

Some studies indicate that the most characteristic profile of people who self-harm are adolescents of both genders and young and adult women (Pérez-Elizondo, 2020), being less likely in adulthood in general. A study by Swannell et al. (2014) observed that the age at which most NSSI appeared was adolescence, between 10 and 17 years (17.2%), followed by young people, up to 24 years (13.4%) and being much lowest in adulthood (5.5%). The age of onset has been stipulated between 13 and 16 years (Muehlenkamp et al., 2018), although this information will vary depending on the study. Specifically, between 7.5 and 42.5% of adolescents between 12 and 14 years old have had some type of self-harming behavior at some point in their lives (Cipriano et al., 2017). The short-term consequences of self-harm can affect social and interpersonal relationships, leading to isolation, feelings of misunderstanding, worry, and discomfort within social and family environments. In the long term, it can become an addiction due to the inability to cope with pain and suffering. Furthermore, this self-harming behavior, when repeated over time, carries the risk of the person developing suicidal ideation.

NSSI and Twitter

The act of sharing experiences of self-harm and suicidal acts in adolescents has grown in recent years on social networks. Abi-Jaoude et al., (2020) observed a 15% growth in posts of stressful situations on social networks between 2013 and 2017. In addition, between 70% of adolescents between 13 and 17 years old own a smartphone and are registered in one or more social networks, which makes it easier to publish and have access to NSSI posts. It was estimated that most of them spend at least 5 hours a day reviewing these networks (Abi-Jaoude et al., 2020). Additionally, social media use has been associated with other problems, such as a higher level of concern about body image and eating disorders (Holland & Tiggeermann, 2017). Other studies have analyzed whether publishing self-harming acts could be potentially dangerous for the psychological health of the individual, becoming a toxic echo chamber that normalizes self-harming behavior among young people and promotes self-harming behavior (Dyson et al., 2016; Lerman et al., 2023; Rowe et al., 2014). It was noted that explicit content, both verbal and visual, rarely appears with a "sensitive material" notice. Furthermore, in most cases, users positively reinforce these publications with comments or "likes" due to the intensity or type of damage and the fact of having published and shared it. In this way, self-harm is normalized, a group of people who share the same activity is reinforced where the act of self-harm and the emotional pain of self-infliction are romanticized (Khasawneh et al., 2021), increasing the future probability of self-harm in order to publish it and maintain this level of social reinforcement.

Twitter, now called X, is a network used around the world with a total number of active users of more than 556 million (Statista, 2023). Young people use this network to share their thoughts and opinions with their peer group, with the purpose of being understood and not being judged for their behaviors (Wang et al., 2017). They go to this social network or similar ones because outside of them they do not know how or who to ask for help or they even believe that asking for help could cause them more problems (Houghton & Joinson, 2012). On Twitter, users can publish texts, images, and videos and share them with other users who can interact with them through comments, retweet content that they find interesting or "like" it. There has been a great deal of interest from researchers in knowing how young people interact with each other on social networks (Alhassan et al., 2021; Khasawneh et al., 2020) and how social networks act in relation to content related to NSSI and their control mechanisms of posts, potentially increasing the likelihood that someone will engage in these behaviors (Baer et al., 2020).

Twitter content is usually textual content, photography and videos using hashtags such as "#blithe" (NSSI), "#Deb" (depression), "#ouchtwt" or "#secretsociety123", so as not to be recognized by the platform (Brown et al., 2018; Moreno et al., 2016). Twitter has a "suicide and self-harm policy" (Twitter, 2023b), they both have control over the content posted about self-harm and offer the possibility for other users to report this content for elimination. This procedure does not prevent the posts from reaching many users and there being a lot of content on Twitter about self-harm, as we will analyze later in this research. They also implement automatic content monitoring, showing users warnings of "potentially sensitive content", although without blocking them.

The literature regarding the verification of content on Twitter usually has different attitudes.

A very positive vision about the control of the content of the platforms and, specifically Twitter, is confirmed by the research by Jhaver et al. (2021). They analyze on this platform the accounts of three influential and extremist people, the research found that Twitter eliminated a very high number of conversations from the platform, and, with this, the toxicity levels of the followers were significantly reduced. The platforms combine moderators who work with them to detect inappropriate behavior (Jhaver et al., 2019; Seering, 2020) and automated tools that eliminate inappropriate messages (Kumar et al., 2021).

The research by Ali et al. (2021), revealed that when people are banned for spreading toxic content on networks, not only does the influence of the banned people decrease, but also the levels of activity and toxicity of their followers. So they conclude that "veto" could be an effective strategy to detoxify social networks. Another study like that, by Naslund et al. (2020), set to find out if users' mood and emotional states could be detected and predicted through the content and images they publish or follow. This could be possible by combining manual content controls with application monitoring in networks. They consider that networks could be beneficial to share their experiences with the disease, seek support from other people and seek information.

A more pessimistic view is found in the work of Lerman et al. (2023). They have a very negative position in relation to Twitter policies, which, although the policies are clear in not tolerating self-harm, the truth is that the hashtag "shtwt" that describes the self-harm subculture on Twitter, has increased by 500% between October 2021 and August 2022 (Goldenberg et al., 2022), with 20,000 tweets per month on average. This indicates that content moderation is not working and is not capable of moderating self-harm content.

In addition, the study by Khasawneh et al. (2021) the objective of their work is to analyze the motivation that leads users to publish content about the three most viral challenges on social networks. One of the criticisms they indicate is the lack of effectiveness in the identification and publication of self-harm content or challenges on video and content sharing platforms. Along these lines is the work of Emma Hilton (2017) who analyzed 362 messages and among the most relevant results, the researcher indicates that self-harming behavior is not understood and is a source of ridicule, which can contribute to delaying access to treatment. The author suggests that Twitter is a medium that can contribute to the healing process by telling personal stories with messages of hope and recovery. But on the other hand, there are also messages that are not very supportive of these behaviors that may not help in the recovery process.

Therefore, our objective is to know how this phenomenon is displayed on social networks and what mechanism the Twitter platform uses to manage this content that, at the same time, is not allowed by the social network's own policies ("relative policy to suicide and self-harm" [Twitter, 2023]) and try to focus on moderation policies so that they are more appropriate and avoid content that may promote self-harm.

OBJECTIVES AND HYPOTHESES

The objectives of this study were two:

1. Analyze the hashtags related to NSSI on Twitter.
2. Analyze the criteria that Twitter follows to veto or block hashtags.

The hypotheses proposed are the following:

1. There is no agreement between Twitter policy and the banned hashtags in relation to NSSI. That is, the regulatory criteria used are not consistent with the published regulatory criteria.
2. There is no logic between banned and unbanned hashtags, they seem to be banned at random. That is, the regulatory criteria used are not consistent among the published material.
3. There is sensitive material that promotes self-harming behavior that is not vetoed.

METHODOLOGY

This work was carried out in two phases through a quantitative study and a qualitative study. In the first study, all tweets containing the words "shtwt", "selfharm", "ouchitweet" or "goretweet" between 11/24/2022 and 01/09/2023 were collected and analyzed. These words were chosen because they are the labels most used by young people with explicit content related to self-harm and with a high number of interactions (Alhassan et al. 2021; Emma Hilton, 2017; Martínez-Pastor et al., 2023; Moreno et al., 2016). A total of 32,231 tweets were collected and those that had been blocked after a month were analyzed. Twitter policies from the year 2023 prior to the change to X were analyzed, it is possible that they may have been changed today.

In the second study, a sample of 100 random tweets was collected and analyzed based on whether they shared text, photo, video, or a combination of the above, whether they were explicit about self-harm, whether they encouraged self-harm or promoted it, and their sensitivity criteria–defined by Twitter–as well as whether or not they had been blocked after three weeks.

Procedure

Study 1

To carry out the first study, a Python application was created that daily downloaded all tweets that contained the words "shtwt", "selfharm", "ouchitweet" or "goretweet". This application made use of Twitter's application programming interface (API), retrieving tweets every 8 hours using the following query: "(shtwt OR selfharm OR ouchietwt OR goretwt)–is: retweet". It was decided to eliminate the retweets of the query, since they did not provide new content.

The result of the queries was stored in a MySQL database, which made it easy to perform complex and aggregate queries that cannot be obtained directly by the Twitter API. Queries of the type: "give me the n tweets with the most impressions between two specific dates" or "what are the n authors with the greatest sum of "likes" in their tweets", can be made on the database with queries in SQL language, but not directly with the Twitter API. Further, the Twitter API limits its results to the last month.

After the tweets were recovered and inserted into the database, it was necessary to periodically update the statistics (number of likes, retweets, and impressions), since these vary permanently. This task becomes difficult if the data set becomes too large for the API limits. First, the tweets were sorted by the time between the creation of the tweet and its last update, and then the list was updated until the API limit was exceeded. This procedure was repeated every hour when access to the API had been restored. When updating the statistics, some tweets began to produce a "tweet not found" error, indicating that they had been blocked or suspended.

Study 2

The first 100 tweets on 7 May 2023, were chosen. They were analyzed if they contained an image, video, or text, if the content was explicit about self-harm, if the self-harm was recent (bleeding) or old (healing) and if it was categorized as sensitive or not by Twitter, in addition to whether the sensitivity of the tweet had an ADD (the Twitter API returned for each search result the attribute "possibly sensitive" with value True or False). In addition, it was specified whether it had been blocked after three weeks and whether it was considered correct or not correct, following the blocking and sensitivity policies published by Twitter (2023b). Twitter defines as "sensitive" any video or image that shows nudity, violent sexual behavior, or is not suitable for minors, according to its document "Policy Regarding Sensitive Multimedia Content" by Twitter (2023a). Also deemed unacceptable are "(...) some types of sensitive media content that may normalize violence and cause

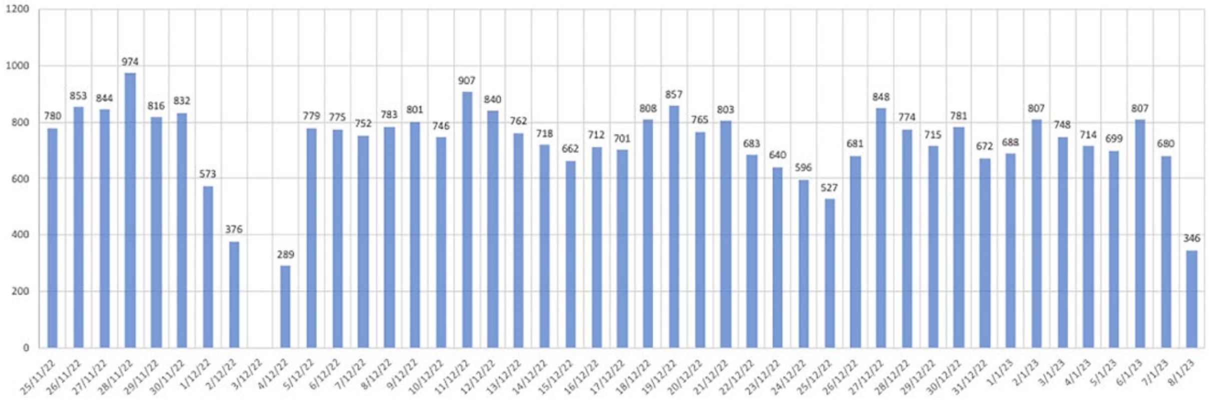

Figure 1. Number of tweets per day (Source: Authors)

suffering to those who view it," including blood, visible wounds, and other behaviors such as self-harm or suicidal thoughts (Twitter, 2023b). They exercise control over content related to self-harm and provide the option for other users to report such content for removal. Additionally, they specify the consequences of not adhering to these rules, explicitly stating that it could result in "blocking access to your account for a period of time before allowing you to tweet again."

RESULTS

In the first study, 32,231 tweets were collected between 11/24/2022 and 01/09/2023 with the following distribution by keywords: "shtwt" (27611; 85.6%); "selfharm" (3342; 10.3%); "ouchietwt" (3110; 9.6%); and "goretwt" (2419; 7.5%). As you can see, "shtwt" (short for "selfharm tweet") is by far the most used keyword. Although these words are used as labels or hashtags, the hashtag is rarely used. Only 6,227 tweets contain '#' (19.3% of the total). This is probably because since they are not words in common language and do not appear in contexts other than self-harm, the hashtag loses its usefulness.

The sample data analyzed shows that a user receives an average of 700 impressions per day (**Figure 1**). In addition, at least 15% of users who wrote in "shtwt" made explicit requests for friendship, or for contacts to interact with. Of the total sample analyzed using the API, 11,361 tweets (35.25%) were automatically identified by Twitter as possibly sensitive. None were preemptively blocked or manually monitored before publication, as it would not have been returned to the API in that case. Of these, some tweets were subsequently blocked, and some users' accounts were even suspended.

Finally, the data of the sample carried out by the API was searched for the word "edtwt", which is the abbreviation of "eating disorder tweet", and 6,069 tweets containing this word were recovered (18.2% of the total), almost 20 percent.

In the second study, the first 97 tweets from 7 May 2023 were chosen to analyze, one by one, whether they had been blocked three weeks later, following the criteria described by Twitter regarding its publication policy. **Figure 2** presents the data from the 97 tweets depending on whether text, image or video was published, whether it referred to self-harm or something else and whether the classification of "sensitive", "non-sensitive" and the decision to block or not block, follow Twitter's criteria.

As can be seen, 54% of the tweets were classified as potentially sensitive compared to 46%. However, the tweets that were finally blocked (35% of the total) are distributed in a similar way between the sensitive and non-sensitive ones (20% of the sensitive ones compared to 15% of the non-sensitive ones), which indicates the limited usefulness or veracity of the classification.

As can be seen, 47.42% (46 tweets) were classified as NOT sensitive. Of these, 100% corresponded to text tweets, without photographs or images. Of these 46 tweets, 10 of them referred to minors self-harming or warning that they were going to do so throughout the day and/or other self-harm groups were proposed so as not to be "caught", etc.

In fact, of the total sample analyzed, using the API, 11,361 tweets (35.25%) were automatically identified by Twitter as possibly sensitive. None were preemptively blocked or manually monitored before publication,

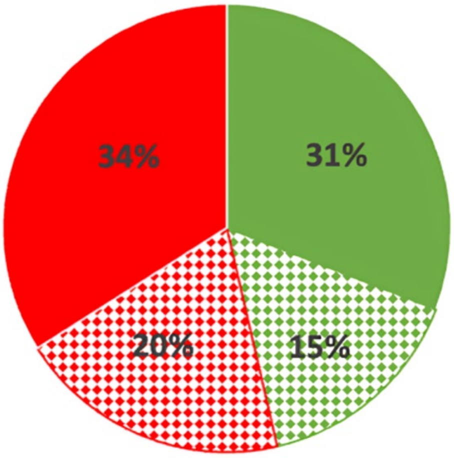

Figure 2. Sensitive blocked (🟥), non-sensitive non-blocked (🟩), non-sensitive blocked (□), and sensitive non-blocked tweets (■) (Source: Authors)

as it would not have been returned to the API in that case. 51 tweets (52.58%) were classified as “sensitive”, of which 45 (88.23%) met the characteristics of being explicit regarding images or videos related to self-harm. The other 11.76% (6 tweets), although they were accompanied by images, were not sensitive and/or did not make explicit reference to self-harm. Two of them referenced a lecture on self-harm, one a plate of beans (“beNSSI” means cutting a specific layer of skin), a photo of a group of girls with sunglasses on and another of a boy with glasses and a microphone, a music group and the logo of a well-known soft drink, which is shaped like blood or a scratch. 100% of the analyzed tweets, which contain a photograph or video, have been classified as “sensitive”, regardless of their content. In the same way, 100% of the videos can only be seen if a click is made, accepting that it is understood to have sensitive content, even if it is not later. However, the photographs, although they have all been classified as sensitive, do not have this prior step accepting possible sensitivity and can be viewed directly, without giving the viewer a choice.

Regarding those who were classified as not sensitive, there were 46. Of these, 78.26% (36 tweets) did not promote self-harm, were not minors and were not visually sensitive. So, they had been classified correctly. However, 27.74% (10 tweets) did explicitly promote self-harm or suicide or were published by minors.

When the same tweets were analyzed again, 3 weeks later (1 June 2023) after having collected them, only 17 tweets (33.33%), of those classified as sensitive, had been blocked.

Finally, 18.2% of the tweets analyzed also referred to some disorder or difficulty related to eating orders.

DISCUSSION

In both studies, almost 20% of the publications referred to NSSI and a problem related to an eating disorder or dissatisfaction with the physical appearance of the body in relation to weight. This reflects the fact that a significant portion of young people who engage in self-harming behavior also have eating problems. Although not all people who suffer from anorexia commit self-harm, as pointed out by Calvete Zumalde et al. (2015). There are many authors who defend that self-injurious behavior occurs more frequently in people who have different types of mental disorders, with eating disorders being the most frequent (Buelens et al., 2020).

In our first hypothesis, we argued that we would find no relationship between the policies proposed by Twitter and blocked tweets. In this sense, Twitter allows the dissemination of this inappropriate content because it does not have preventive control given that, as stated above, of the total sample only 35.25% were considered possibly sensitive by the Twitter platform, which could normalize self-harm and encourage these behaviors as highlighted by Lookingbill (2022) and Emma Hilton (2017). Twitter defines as “sensitive” any video or image that depicts nudity and violent sexual behavior or is not suitable for minors, in its document “Policy Regarding Sensitive Multimedia Content” (Twitter, 2023a). Later, comment on other content that would be equally unacceptable “(…) some types of sensitive multimedia content that we do not allow under any circumstances, because they could normalize violence and cause suffering to those who view them” (Twitter, 2023a). These other forms of multimedia content refer to violent crimes or accidents; physical fights; physical abuse of minors; bodily fluids, including blood, feces, semen, etc.; serious physical damage, such as visible wounds; and seriously injured or mutilated animals. Additionally, in its policy related to suicide and self-harm, as has been observed throughout this article, these measures are not met, since a large number of tweets refer to the explicit visualization and/or verbal promotion of self-harm. In relation, for example, to the "gratuitous bloody scenes", where they could rate the majority of the selected tweets, following their policy on sensitive multimedia content, they comment that, if this type of material is shared, the user will be asked to delete it, whether they are videos or images. After the first warning, they may suspend the account permanently. In relation to self-harm, where the act or the consequence of the same is graphically shown, it should always be eliminated, since, according to this same policy, Twitter is only allowed to publish videos and images related to sexuality where the consent of the people involved is evident and warn that it is sensitive viewing ("policy regarding suicide and self-harm with the aim of preventing these behaviors"). Despite these warnings, three weeks later, only 33% of the images classified as "sensitive" and that violated both policies were blocked. None of these images or texts contribute to a better understanding of the mechanisms of self-harm, reflection on them, how to avoid them, or seek help, except for 2 tweets that make explicit reference to a conference on self-harm. These two tweets, therefore, have been automatically classified as "sensitive" due to the fact that they carry an image. These data agree with those found by the Network Contagion Research Institute (Goldenberg et al., 2022), in which they reported that the use of groups related to self-harm on Twitter had increased by 500% and that posts contrary to their policy related to suicide and self-harm.

Videos, photographs and even texts related to self-harm develop habituation in the viewer. Habituation is a phenomenon, based on the laws of learning, that explains how people react less and less to habitual stimuli, although these stimuli increase in intensity. In the same way, for example, the insensitivity that many people have towards violence is explained, by exposing it publicly and explicitly in both a narrative and visual way. Funkhouser et al. (2019) conducted a study on response time to aversive stimuli, comparing women who had practiced self-harm and a control group.

In many cases, furthermore, it not only gives rise to "normalization" but can be conditioned as something positive or pleasant, for example, when it is accompanied by humorous comments that generate laughter or with admirable attributes of the person, such as courage. Further increasing the probability of consumption and imitation of said material (Memon et al., 2018). In the suicide and self-harm policy, in fact, Twitter itself confirms that any positive comments or comments that encourage self-harming behavior are prohibited. In the study by Lewis and Seko (2016), they also observed that, although these groups could have some positive aspects, such as feeling like they belonged to a group of equals, it had many negative consequences. Specifically, the exposure increased the urge to self-harm and the post reinforced self-injurious behavior. In our sample, comments have been found where self-harm is normalized as part of daily life. For example, a photo of recent cuts appears and as a comment, the user writes, "I'm going to pee and have a twinkie." However, these tweets have not been blocked, even if they are accompanied by images of bleeding wounds. Manual review of the tweets confirms that the majority of photographs show open cuts, blood and/or blades. Some photographs are extremely violent and many also trivialize self-injury, talking about it as something every day or a daily task and infantilizing it. Other studies have observed that people who self-harm as a form of emotional regulation have lower levels of cognitive maturity and higher levels of impulsivity (Hjelmeland & Grøholt, 2005). Thus, in our sample you can see kitten-shaped blades or manga figures in attractive and erotic postures, wearing school clothes, and with self-inflicted injuries. Furthermore, there are studies that indicate that exposure to what is "forbidden" leads to an increase in adrenaline and norepinephrine as a form of alert, which does not allow us to look away (Chappell, 2022). This is used by many media outlets to attract the audience. This generates repeated exposure, habituation, imitation and the cycle explained above.

Our second hypothesis argued that we would not find consistency between tweets that have been blocked and those that have not been blocked after 3 weeks. Paying attention to the texts, for example, we have observed in the different examples of the second study that X does not consider any text "sensitive", regardless of the content. This includes content with suicidal ideation and planning, description of self-harm or self-harm intention, etc., including those that come from people who acknowledge being minors. To block, they only take images into account, not text. To justify the sensitivity, they also do not take the text into account. For example, in three of the tweets there was a photo about a conference on self-harm. The photos were harmless and yet they are described as sensitive. In another tweet, however, only text appears in which the user specifically promotes membership in a group that cuts off body parts and cooks them. As seen, there is no coherence between Twitter's criteria and the tweets that are ultimately blocked or announced as sensitive.

Contrary to its own policy, which includes both videos and photographic material as "sensitive", no photographs have been found that warn that sensitive material is going to be presented and allow the consumer to decide whether they want to open it or not. In the case of videos, however, it exists 100% of the time. The fact that the "sensitivity" rules are only applied to videos, but not to images or written messages, seems flawed and appears to be a decision made by an algorithm rather than a decision made by detailed analysis of the content of the tweets.

According to the **third hypothesis,** there is sensitive material that promotes self-injurious behavior that is not off-limits. No criterion has been found that justifies which images overlap and which do not. Although all the photos have been classified as sensitive, which means that the only existing filter is the one that identifies that there is a photographic file, but its content is not analyzed, it is an automatic filter that does not include a detailed review of the tweets related to self-harm. For example, after three weeks, minor self-harm had disappeared, with superficial cuts or even bandaged wounds where the self-harm was not visible, but there was still very severe self-harm with deep cuts, bleeding and showing various layers of the skin. On the other hand, in 4 tweets they present themselves as new users/consumers. The only one classified as "sensitive" is the one in which a photo has been attached, even though the other three identify themselves as people who self-harm and two of them (50% verbalize in said text that they are minors). Those tweets that do not have a photo are considered low sensitivity, regardless of the textual content, even though some of the verbalizations express sadistic feelings or self-harming or suicidal intentions.

Following Twitter's policy, in addition, minors should not, under any circumstances, follow content referring to self-harm and, therefore, be part of these groups. However, there are multiple examples found of minors who publish and minors who follow. Both texts, e.g., "I'm cutting myself in the school bathroom," and images. In a tweet, a user says "minors DNI." Twitter requires that the user be at least 13 years old but does not have any system that allows it to verify the age of said user.

Considering the examples that have been presented throughout the discussion, we observe that the majority of the tweets analyzed in the second study show, reinforce, and promote self-harming behavior. Their messages have comments that generate laughter or admiration due to attributes such as courage. There are also images of bloody wounds that stimulate their behavior, further increasing the likelihood of consumption and imitation. In many comments, self-harm is normalized as part of daily life. The solutions that different authors give to this lack of homogeneous control are several: 1) combine exhaustive manual controls of the content with monitoring through applications (Jhaver et al., 2021; Lerman et al., 2023); 2) elimination of groups on platforms based on toxicity levels in the networks (Ali et al., 2021); 3) create scales that indicate the level of harmfulness of the content of a publication (Khasawneh et al., 2021). Still others indicate that social media could provide valuable data to support effective public education campaigns and personalized and supportive treatment options and for young people to share their lived experiences with the disease (Emma Hilton, 2017; Naslund et al. 2020), but, from our analysis, always with clear and adequate supervision of the content disseminated by the platform.

CONCLUSIONS

The study of tweets of 4 keywords related to self-harm has shown that a person who follows these keywords will receive a daily average of 725 daily impacts, the vast majority of which are images of great violence. The impact that these violent images have on young people is very harmful to their mental health, may cause the normalization, imitation, and romanticization of self-harm. It also produces in adolescents, who regularly consume this content, the feeling of belonging to a group, which reinforces the phenomenon of self-harm.

The application of Twitter's policies does not appear to be effective in view of our results. Additionally, they create in adolescents who habitually consume this content a sense of belonging to a group, which reinforces the phenomenon of self-harm. Although some tweets and users are blocked, this blocking is done when they have already produced a high number of impacts on other users. Automatic classification of content as possibly sensitive does not adequately distinguish content. Additionally, content that Twitter identifies as sensitive is not blocked preventively for minors, which seems irresponsible when, despite being aware of the danger - as derived from the fact of having policies in this regard - they do not implement any age controls either.

Author contributions: DA-M, CN-G, & EM-P: bibliographic search, investigation, writing–original draft, writing–review & editing, formal review; **DA-M & CN-G:** methodology, statistics, validation of calculations; **DA-M & EM-P:** conceptualization, project administration, **DA-M:** materials & resources, software, visualization & graphics, curation of the used data, supervision; **EM-P:** fund acquisition. All authors approved the final version of the article.

Funding: This article was funded by the projects "Media representation of minors' self-harm in the media and networks" (PID2021-124550OB-I00) by the National Plan of the Ministry of Science and Innovation and SIC-SPAIN 4.0 (UE).

Ethics declaration: This study uses previously published data and does not involve personal data or direct contact with individuals. The research was reviewed and approved by the Ethics Committee of the Universidad Rey Juan Carlos (n. approval: 0802202307023).

Declaration of interest: The authors declare no competing interest.

Data availability: Data generated or analyzed during this study are available from the authors on request.

REFERENCES

Abi-Jaoude, E., Naylor, K. T., & Pignatiello, A. (2020). Smartphones, social media use and youth mental health. *Canadian Medical Association Journal, 192*(6), E136–E141. https://doi.org/10.1503/cmaj.190434

Alhassan, M. A., Inuwa-Dutse, I., Bello, B. S., & Pennington, D. (2021). Self-harm: Detection and support on Twitter. In *Proceedings of the 8th European Conference on Social Media*. Academic Conferences International. https://doi.org/10.48550/arXiv.2104.00174

Ali, S., Saeed, M. H., Aldreabi, E., Blackburn, J., de Cristofaro, E., Zannettou, S., & Stringhini,G. (2021). Understanding the effect of deplatforming on social networks. In *Proceedings of the 13th ACM Web Science Conference* (pp. 187–195). ACM. https://doi.org/10.1145/3447535.3462637

American Psychiatric Association. (2014). *Diagnostic and statistical manual of mental disorders*. American Psychiatric Association Publishing. https://doi.org/10.1176/appi.books.9780890425596

American Psychiatric Association. (2022). *Diagnostic and statistical manual of mental disorders*. American Psychiatric Association Publishing. https://doi.org/10.1176/appi.books.9780890425787

Angelakis, I., & Gooding, P. (2021). Experiential avoidance in non-suicidal self-injury and suicide experiences: A systematic review and meta-analysis. *Suicide and Life-Threatening Behavior, 51*(5), 978–992. https://doi.org/10.1111/sltb.12784

Baer, M. M., Tull, M. T., Forbes, C. N., Richmond, J. R., & Gratz, K. L. (2020). Methods matter: Nonsuicidal self-injury in the form of cutting is uniquely associated with suicide attempt severity in patients with substance use disorders. *Suicide and Life-Threatening Behavior, 50*(2), 397–407. https://doi.org/10.1111/Sltb.12596

Barrocas, A. L., Hankin, B. L., Young, J. F., & Abela, J. R. (2012). Rates of nonsuicidal self-injury in youth: Age, sex, and behavioral methods in a community sample. *Pediatrics, 130*(1), 39–45. https://doi.org/10.1542/peds.2011-2094

Brown, R. C., Fischer, T., Goldwich, A. D., Keller, F., Young, R., & Plener, P. L. (2018). # cutting: Non-suicidal self-injury (NSSI) on Instagram. *Psychological Medicine, 48*(2), 337–346. https://doi.org/10.1017/S0033291717001751

Buelens, T., Luyckx, K., Verschueren, M., Schoevaerts, K., Dierckx, E., Depestele, L., & Claes, L. (2020). Temperament and character traits of female eating disorder patients with (out) non-suicidal self-injury. *Journal of Clinical Medicine, 9*(4), Article 1207. https://doi.org/10.3390/jcm9041207

Calvete Zumalde, E., Orue Sola, I., Aizpuru, L., & Brotherton, H. (2015). Prevalence and functions of non-suicidal self-injury in Spanish adolescents. *Psicothema, 27*(3), 223–228. https://doi.org/10.7334/psicothema2014.262

Chappell, Z. (2022). The enacted ethics of self-injury. *Topoi, 41*, 383–394. https://doi.org/10.1007/s11245-022-09796-z

Cipriano, A., Cella, S., & Cotrufo, P. (2017). Nonsuicidal self-injury: A systematic review. *Frontiers in Psychology, 8*, Article 1946. https://doi.org/10.3389/fpsyg.2017.01946

Dyson, M. P., Hartling, L., Shulhan, J., Chisholm, A., Milne, A., Sundar, P., Scott, S. D., & Newton, A. S. (2016). A systematic review of social media use to discuss and view deliberate self-harm acts. *PLoS ONE, 11*(5), Article e0155813. https://doi.org/10.1371/journal.pone.0155813

Emma Hilton, C. (2017). Unveiling self-harm behaviour: What can social media site Twitter tell us about self-harm? A qualitative exploration. *Journal of Clinical Nursing, 26*(11–12), 1690–1704. https://doi.org/10.1111/jocn.13575

Faura-García, J., Calvete Zumalde, E., & Orue Sola, I. (2021). Autolesión no suicida: Conceptualización y evaluación clínica en población hispanoparlante [Non-suicidal self-harm: Conceptualization and clinical evaluation in the Spanish-speaking population]. *Papeles del Psicólogo, 42*(3), 207–214. https://doi.org/10.23923/pap.psicol.2964

Favazza, A. R., & Favazza, B. (1987). *Bodies under siege: Self-mutilation in culture and psychiatry*. Johns Hopkins University Press.

Funkhouser, C. J., Correa, K. A., Carrillo, v. L., Klemballa, D. M., & Shankman, S. A. (2019). The time course of responding to aversiveness in females with a history of non-suicidal self-injury. *International Journal of Psychophysiology, 141*, 1–8. https://doi.org/10.1016/j.ijpsycho.2019.04.008

Goldenberg, A., Farmer, J., Jussim, L., Sutton, L., Finkelstein, D., Ramos, C., Paresky, P., & Finkelstein, J. (2022). *Online communities of adolescents and young adults celebrating, glorifying, and encouraging self-harm and suicide are growing rapidly on Twitter*. https://networkcontagion.us/

Hjelmeland, H., & Grøholt, B. (2005). A comparative study of young and adult deliberate self-harm patients. *Crisis, 26*(2), 64–72. https://doi.org/10.1027/0227-5910.26.2.64

Holland, G., & Tiggemann, M. (2017). "Strong beats skinny every time": Disordered eating and compulsive exercise in women who post fitspiration on Instagram. *International Journal of Eating Disorders, 50*(1), 76–79. https://doi.org/10.1002/eat.22559

Houghton, D. J., & Joinson, A. N. (2012). Linguistic markers of secrets and sensitive self-disclosure in Twitter. In *Proceedings of the 45th Hawaii International Conference on System Sciences* (pp. 3480–3489). IEEE. https://doi.org/10.1109/HICSS.2012.415

Jhaver, S., Birman, I., Gilbert, E., & Bruckman, A. (2019). Human-machine collaboration for content regulation: The case of reddit automoderator. *ACM Transactions on Computer-Human Interaction, 26*(5), Article 31. https://doi.org/10.1145/3338243

Jhaver, S., Boylston, C., Yang, D., & Bruckman, A. (2021). Evaluating the effectiveness of deplatforming as a moderation strategy on Twitter. *Proceedings of the ACM on Human-Computer Interaction, 5*(CSCW2), Article 381. https://doi.org/10.1145/3479525

Khasawneh, A., Madathil, K. C., Dixon, E., WiśNiewski, P., Zinzow, H., & Roth, R. (2020). Examining the self-harm and suicide contagion effects of the Blue Whale Challenge on YouTube and Twitter: Qualitative study. *JMIR Mental Health, 7*(6), Article e15973. https://doi.org/10.2196/15973

Khasawneh, A., Madathil, K. C., Zinzow, H., Wisniewski, P., Ponathil, A., Rogers, H., Agnisarman, S., Roth, R., & Narasimhan, M. (2021). An investigation of the portrayal of social media challenges on YouTube and Twitter. *ACM Transactions on Social Computing, 4*(1), Article 2. https://doi.org/10.1145/3444961

Klonsky, E. D. (2011). Non-suicidal self-injury in United States adults: Prevalence, sociodemographics, topography and functions. *Psychological Medicine, 41*(9), 1981–1986. https://doi.org/10.1017/S0033291710002497

Klonsky, E. D., & Olino, T. M. (2008). Identifying clinically distinct subgroups of self-injurers among young adults: A latent class analysis. *Journal of Consulting and Clinical Psychology, 76*(1), Article 22. https://doi.org/10.1037/0022-006X.76.1.22

Kumar, D., Kelley, P. G., Consolvo, S., Mason, J., Bursztein, E., Durumeric, Z., Thomas, K., & Bailey, M. (2021). Designing toxic content classification for a diversity of perspectives. In *Proceedings of the 17th Symposium on Usable Privacy and Security* (pp. 299–318). USENIX. https://doi.org/10.48550/arXiv.2106.04511

Lerman, K., Karnati, A., Zhou, S., Chen, S., Kumar, S., He, Z., Yau, J., & Horn, A. (2023). Radicalized by thinness: Using a model of radicalization to understand pro-anorexia communities on Twitter. *arXiv*. https://doi.org/10.48550/arXiv.2305.11316

Lewis, S. P., & Seko, Y. (2016). A double-edged sword: A review of benefits and risks of online nonsuicidal self-injury activities. *Journal of Clinical Psychology, 72*(3), 249–262. https://doi.org/10.1002/jclp.22242

Lookingbill, V. (2022). Examining nonsuicidal self-injury content creation on TikTok through qualitative content analysis. *Library & Information Science Research, 44*(4), Article 101199. https://doi.org/10.1016/j.lisr.2022.101199

Martínez-Pastor, E., Atauri-Mezquida, D., Nicolás-Ojeda, M. Á., & Blanco-Ruiz, M. (2023). Visualización e interpretación de las interacciones en los mensajes de autolesiones no suicidas (NSSI) en Twitter [Visualizing and interpreting interactions on non-suicidal self-injury (NSSI) messages on Twitter]. *Revista Hispana Para el Análisis de Redes Sociales, 34*(2), 238–253. https://doi.org/10.5565/rev/redes.996

Memon, A. M., Sharma, S. G., Mohite, S. S., & Jain, S. (2018). The role of online social networking on deliberate self-harm and suicidality in adolescents: A systematized review of literature. *Indian Journal of Psychiatry, 60*(4), Article 384. https://doi.org/10.4103/psychiatry.IndianJPsychiatry_414_17

Moreno, M. A., Ton, A., Selkie, E., & Evnssi, Y. (2016). Secret society 123: Understanding the language of self-harm on Instagram. *Journal of Adolescent Health, 58*(1), 78–84. https://doi.org/10.1016/j.jadohealth.2015.09.015

Muehlenkamp, J. J., Xhunga, N., & Brausch, A. M. (2018). Self-injury age of onset: A risk factor for NSSI severity and suicidal behavior. *Archives of Suicide Research, 23*(4), 551–563. https://doi.org/10.1080/13811118.2018.1486252

Naslund, J. A., Bondre, A., Torous, J., & Aschbrenner, K. A. (2020). Social media and mental health: Benefits, risks, and opportunities for research and practice. *Journal of Technology in Behavioral Science, 5*, 245–257. https://doi.org/10.1007/s41347-020-00134-x

Nock, M. K. (2010). Self-injury. *Annual Review of Clinical Psychology, 6*, 339–363. https://doi.org/10.1146/annurev.clinpsy.121208.131258

Nock, M. K., & Favazza, A. R. (2009). Nonsuicidal self-injury: Definition and classification. In M. K. Nock (Ed.), *Understanding nonsuicidal self-injury: Origins, assessment, and treatment* (pp. 9–18). American Psychological Association. https://doi.org/10.1037/11875-001

Pattison, E. M., & Kahan, J. (1983). The deliberate self-harm syndrome. *The American Journal of Psychiatry, 140*(7), 867–872. https://doi.org/10.1176/ajp.140.7.867

Pérez-Elizondo, A. D. (2020). ¿Qué es el síndrome FOMO? [What is FOMO syndrome?] *Psicología. com*. https://psiquiatria.com/trabajos/usr_7775066768657.pdf

Pérez-Elizondo, A. D. (2021). Enfermedad por autolesión. ¡Primero me corto, luego existo [Enfermedad por autolesión. ¡Primero me corto, luego existo]! *Archivos de Investigación Materno Infantil, 11*(2), 77–81. https://doi.org/10.35366/101554

Rowe, S. L., French, R. S., Henderson, C., Ougrin, D., Slade, M., & Moran, P. (2014). Help-seeking behaviour and adolescent self-harm: A systematic review. *Australian & New Zealand Journal of Psychiatry, 48*(12), 1083–1095. https://doi.org/10.1177/0004867414555718

Seering, J. (2020). Reconsidering self-moderation: The role of research in supporting community-based models for online content moderation. *Proceedings of the ACM on Human-Computer Interaction, 4*(CSCW2), Article 107. https://doi.org/10.1145/3415178

Stacy, S. E., Bandel, S. L., Lear, M. K., & Pepper, C. M. (2018). Before, during, and after self injury: The practice patterns of nonsuicidal self-injury. *The Journal of Nervous and Mental Disease, 206*(7), 522–527. https://doi.org/10.1097/NMD.0000000000000846

Staniland, L., Hasking, P., Boyes, M., & Lewis, S. (2021). Stigma and nonsuicidal self-injury: Application of a conceptual framework. *Stigma and Health, 6*(3), Article 312. https://doi.org/10.1037/sah0000257

Statista. (2023). *Redes sociales con mayor número de usuarios activos a nivel mundial en enero de 2022 (en millones)* [Social networks with the highest number of active users worldwide in January 2022 (in millions)]. https://es.statista.com/estadisticas/600712/ranking-mundial-de-redes-sociales-pornumero-de-usuarios

Sutton, J. (2007). *Healing the hurt within: Understand self-injury and self-harm, and heal the emotional wounds*. How To Books.

Swannell, S. V., Martin, G. E., Page, A., Hasking, P., & St John, N. J. (2014). Prevalence of nonsuicidal self-injury in nonclinical samples: Systematic review, meta-analysis and meta-regression. *Suicide and Life-Threatening Behavior, 44*(3), 273–303. https://doi.org/10.1111/sltb.12070

Twitter. (2023a). *Policy regarding sensitive multimedia content*. Retrieved June 2023, from https://help.twitter.com/es/rules-and-policies/media-policy

Twitter. (2023b). *Suicide and self-harm policy*. Retrieved June 2023, from https://help.twitter.com/en/rules-and-policies/glorifying-self-harm

UNICEF. (2020). *Annual report 2020*. https://www.unicef.org/reports/unicef-annual-report-2020

UNICEF. (2021a). *Por lo menos 1 de cada 7 niños y jóvenes ha vivido confinado en el hogar durante gran parte del año, lo que supone un riesgo para su salud mental y su bienestar* [At least 1 in 7 children and young people have been confined to their homes for much of the year, posing a risk to their mental health and well-being]. https://www.unicef.org/es/comunicados-prensa/1-cada-7-ninos-jovenes-ha-vivido-confinado-hogar-durante-gran-parte-ano

UNICEF. (2021b). *Estado mundial de la infancia 2021* [State of the world's children 2021]. https://www.unicef.org/es/informes/estado-mundial-de-la-infancia-2021

Wang, Y., Tang, J., Li, J., Li, B., Wan, Y., Mellina, C., O'Hare, N., & Chang, Y. (2017). Understanding and discovering deliberate self-harm content in social media. In *Proceedings of the 26th International Conference on World Wide Web* (pp. 93–102). ACM. https://doi.org/10.1145/3038912.3052555