Article

Social Science Computer Review
2020, Vol. 38(6) 673-685

DOI: 10.1177/0894439319836389
journals.sagepub.com/home/ssc
SAGE

Detecting Intentional Self-Harm on Instagram: Development, Testing, and Validation of an Automatic Image-Recognition Algorithm to Discover Cutting-Related Posts

Sebastian Scherr[1], Florian Arendt[2], Thomas Frissen[1], and José Oramas M. [1]

Abstract

Self-injurious behavior is often practiced in secrecy or involves body parts that are easy to hide, making early detection difficult and hampering intervention and treatment. However, cutting, one of the most common intentional forms of nonsuicidal self-injury (NSSI), is relatively often shared publicly via new digital media technologies. We explored NSSI on Instagram through a pioneering combination of two computational methods: First, we developed an automatic image-recognition algorithm that uncovered NSSI (or the absence of NSSI) in digital pictures, and second, we employed web-scraping techniques to obtain all pictures posted on Instagram in a given time frame under four NSSI-related hashtags in English and German. The image-recognition algorithm was then used to explore the relative prevalence of NSSI in these $N = 13{,}132$ pictures posted within 48 hr on Instagram under #cutting ($n = 4{,}219$), #suicide ($n = 7{,}910$), #selbstmord ($n = 173$), and #ritzen ($n = 830$) in June 2018. This article not only aims to raise awareness of NSSI on Instagram but also introduces the first automatic image-recognition algorithm that addresses cutting on social media and presents that algorithm's first empirical test run on a large sample of pictures scraped from Instagram. The ultimate goal of this research is to protect vulnerable populations from contact with NSSI-related pictures posted on social media.

Keywords

nonsuicidal self-injury (NSSI), intentional self-harm, Instagram, #cutting, automatic image-recognition algorithm

[1] KU Leuven, Leuven, Belgium
[2] University of Vienna, Vienna, Austria

Corresponding Author:
Sebastian Scherr, KU Leuven, Parkstraat 45, Leuven 3000, Belgium.
Email: sebastian.scherr@kuleuven.be

Nonsuicidal self-injury (NSSI) is an umbrella term for a specific form of intentional self-harm without suicidal intent, and unlike tattoos or body piercings, for example, this form of body modification is not socially accepted (De Riggi, Moumne, Heath, & Lewis, 2016). Particularly prevalent among developing youths, the most common forms of NSSI are scratching, bruising, headbanging, burning, and cutting (Nock & Favazza, 2009). Less than a quarter of those who engage in NSSI seek or receive health care before or afterward (Dyson et al., 2016). Importantly, NSSI is linked to individual stress (Hamza, Stewart, & Willoughby, 2012), victimization (O'Connor, Rasmussen, & Hawton, 2009), and exposure to pro-self-harm online content (Minkkinen, Oksanen, Kaakinen, Keipi, & Räsänen, 2017), and both nonsuicidal and suicidal self-harm share a range of risk factors, thus placing NSSI and suicide conceptually on the same continuum (see Mars et al., 2014).

Instagram is the world's second most used social networking site and is particularly popular among young users, with 71% of the site's user base being emerging adults between 18 and 24 years of age (Smith & Anderson, 2018). The platform is primarily visual and driven by the photos that are uploaded onto it. As do Twitter users (Wang, Wei, Liu, Zhou, & Zhang, 2011), Instagram users label uploaded content with key words that are prefixed by the hash sign (e.g., #holidays, #metoo). These key words or hashtags are organically transformed into hyperlinks that redirect the user to all other content on the platform that has been uploaded under that specific hashtag. In that sense, hashtags serve as semantic annotations that allow users to "self-curate"-specific thematic content on Instagram or on other social media platforms (Meraz, 2017).

Since the end of 2017, users can not only search for pictures related to a specific hashtag but can also follow one or more hashtags (Instagram, 2017), which then causes posts under these hashtags to be more likely to show up prominently in the users' Instagram feeds. Some hashtags typically co-occur such as #depression and #cutting, with 27% of the pictures tagged with #depression also being tagged with #cutting (Brown et al., 2018). Importantly, Instagram acknowledges potentially problematic hashtags such as #cutting and shows a content warning message (see Figure 1) that requires users to confirm their consent by clicking before they can access this content. Hashtags can also be shut down completely.

Systematic reviews (Dyson et al., 2016; Lewis & Seko, 2016) have identified the benefits (social support, coping, emotional self-disclosure, and inhibition of NSSI) and risks (reinforcement, triggers, and stigmatization of NSSI) of engaging with NSSI content online with considerable individual differences (Baker & Lewis, 2013). Nock and Prinstein (2004) have suggested that positive social reinforcement may be one reason why exposure to #cutting pictures on Instagram might be dangerous, particularly when such reinforcement manifests as the visible comments or "likes" of others on the social media platform (Brown et al., 2018). However, according to the theory of planned behavior (Ajzen, 1991; Ajzen & Fishbein, 1973), perceived social norms may also be relevant to action. For example, the mere number of pictures subsumed under #cutting depicting people cutting their wrists might shift descriptive norms about the pervasiveness of wrist-cutting as a form of NSSI and thereby trigger future behavioral intentions. In fact, victimization often brings people to NSSI content online (Minkkinen et al., 2017), which further intensifies the relevance of normalization or reinforcing mechanisms on social media. Radovic and Hasking (2013) showed evidence for a modeling and normalization explanation for films. Importantly, NSSI-discouraging messages are often absent on social media such as Instagram (Miguel et al., 2017), thereby intensifying the possible effects. In effect, comments and likes can be regarded as additional social cues and might therefore operate in a reinforcing manner, as noted by Brown et al. (2018). Moreover, research on traditional fictional (Stack, 2009) and nonfictional media (Niederkrotenthaler et al., 2012) has shown that exposure to mediated exemplars can increase suicidal thoughts and behaviors—a phenomenon called the Werther effect (Phillips, 1974). Although evidence for Werther effects in the social media domain is still very limited, it is likely that content posted on social media operates

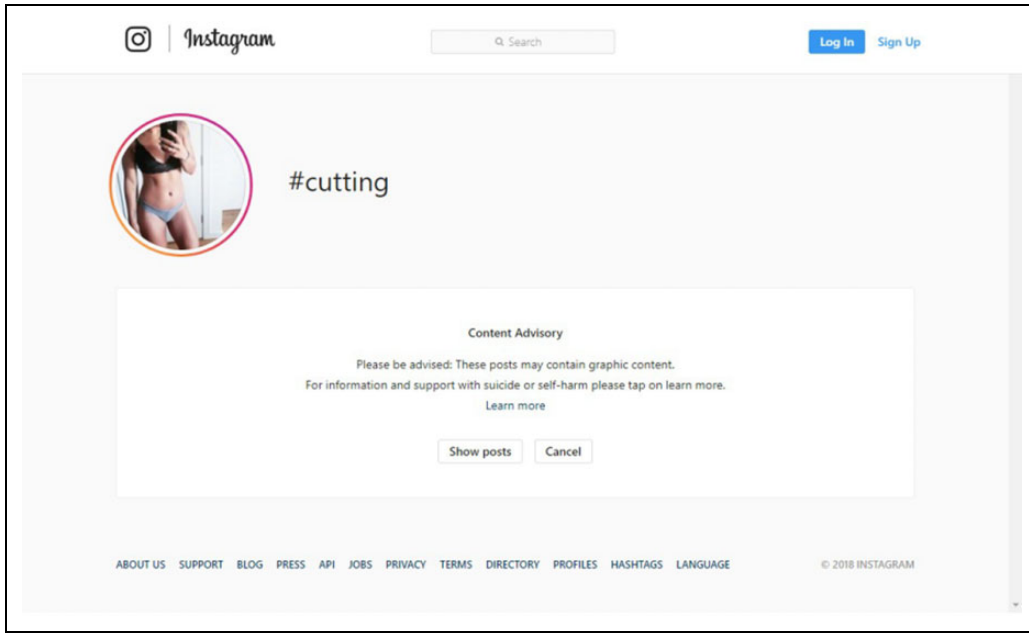

Figure 1. Content warning when accessing #cutting on Instagram.

similarly to other content and has the power to trigger imitational behaviors (see Fahey, Matsubayashi, & Ueda, 2018).

Existing knowledge about the number of pictures related to cutting and NSSI on Instagram is similarly still very limited. Most insights can be drawn from only two recent studies (Brown et al., 2018; Moreno, Ton, Selkie, & Evans, 2016) showing that around two thirds of NSSI pictures posted under #cutting depict healed superficial wounds or scars, while one third depicts fresh wounds. Common English-language hashtags for cutting-related NSSI are #selfharmmm (with one or more extra Ms if reported and shut down), #cat, #selfinjuryyy, or #blithe (Moreno et al., 2016). Currently, the most sophisticated insights about NSSI on Instagram stem from a recent study that explored 30 different cutting-related hashtags in German within a 72-hr time frame (Brown et al., 2018). The study found that slightly more than 5% of all pictures uploaded on Instagram under these hashtags ($n = 293$) contained pictures with wounds or scars (Brown et al., 2018). Over the course of 4 weeks, the upload rate of publicly available cutting-related pictures was 8.8% ($n = 2{,}826$). Within these 4 weeks, 93.1% of the pictures depicted some form of cut (39.6% depicted superficial wounds or scratches, 47.8% showed deeper cuts with blood, and 12.6% were deep, gaping cuts with large amounts of blood). Bruises, bites, burning, and skin picking were less commonly posted. Additionally, 59.6% of the pictures focused on cuts on the upper extremities (e.g., arms, wrists) and the objects that caused the injury (e.g., razorblades, nails, and knives) were seldom depicted (5.8%). In 34 pictures, the cuts into the skin even showed a readable text such as "HATE MYSELF" or "DREAM" (Brown et al., 2018). Temporal patterns were not recognizable for the photographs in Germany, even though slightly more pictures were uploaded on Sundays (Brown et al., 2018). Over the course of the day, more NSSI pictures were posted in the early morning and later in the evening. NSSI postings on Instagram might therefore be related to the stress experienced in the evenings and early mornings before having to go to school (Brown et al., 2018).

What should be clear at this point in our argument is that previous research has already acknowledged the importance of NSSI-related posts on Instagram. One remaining issue, however, is the

sheer amount of content posted on the platform under NSSI-related hashtags and how to measure it. In fact, previous scholarly work has used human coders to categorize Instagram posts. Unfortunately, a huge amount of work is required to code the entirety of the material manually. In addition, when humans code Instagram posts, there is a time lag between identifying, downloading, and coding NSSI content. Therefore, given the conceptual (Mars et al., 2014) and empirical (Brown et al., 2018) overlaps between NSSI and suicide, as well as the link between exposure to online NSSI and increased self-harm and suicidality (Mitchell, Wells, Priebe, & Ybarra, 2014), an automated form of recognition that identifies NSSI-related posts in real time would substantially contribute to NSSI- and suicide-prevention research and practice.

The present article introduces the first automatic image-recognition algorithm that automatically addresses cutting on social media. To overcome the existing limitations, we combine the two computational methods known as web scraping and computer vision: Web-scraping techniques were used to automatically download content posted on Instagram under different problematic hashtags and in different languages, and an image-recognition algorithm was developed to automatically detect NSSI (vs. the absence of NSSI) in Instagram pictures.

As the first study of its kind, this research used an automatic image-recognition algorithm to explore the *prevalence of NSSI (vs. the absence of NSSI) in all pictures* posted under the multilanguage hashtags #cutting, #suicide, #ritzen (= cutting), and #selbstmord (= suicide) on Instagram within a 48-hr time frame in June 2018. In the following sections, we will introduce and describe in more detail both methodological elements of this study. The ultimate goal of this research is the protection of vulnerable populations from contact with NSSI-related pictures posted on social media.

Method

Development of an Automatic Image-Recognition Algorithm

As a first step, we built a database for algorithm training that included a selection of NSSI-related pictures from Instagram. We downloaded two batches of publicly available square-shaped pictures uploaded to the social media platform Instagram. The pictures were downloaded together with a time stamp and the defining hashtags. In accordance with the platform's terms and conditions, no other information about the uploading users, other users' likes, comments on, or shares of the pictures, or the geolocation was downloaded or stored. We used a total of 600 cutting-related images uploaded to the platform between October 2017 and January 2018 that used the hashtags #suizid, #selbstmord, and #ritzen (German for "cutting"). The download was managed using the software program *4K Stogram*. Since the software downloaded all pictures uploaded to Instagram under a specific hashtag, we manually identified pictures that depicted fresh wounds or red scars from cutting body parts.

The training of the machine-learning algorithm also required "negatives"—pictures that did not depict cutting-related content. To include such content, we used a random sample of same-sized images depicting a variety of different content representing negative (i.e., "noncutting") content. Pictures were taken from the ImageNet Large-Scale Visual Recognition Challenge 2012 (ILSVRC '12; Russakovsky et al., 2015). To automatically distinguish depictions of cutting from other images, we trained an automatic image-recognition model. This model was based on a convolutional neural network classifier—more specifically, the AlexNet architecture proposed by Krizhevsky, Sutskever, and Hinton (2012). This architecture achieved state-of-the-art performance for the large-scale 1,000-class classification task from ILSVRC '12. This type of model can learn abstract concepts such as "cutting" from simpler ones such as lines, colors, or shapes. That is, when an image is pushed into

the model, a conclusion concerning a specific task can be reached as a function of the results of intermediate concepts at different internal levels of the model.

The training routine consists of several training cycles that follow the protocol proposed by Krizhevsky et al. (2012) in which two balanced sets of "training" and "testing" images are defined. These sets are composed of image-label pairs (in this study, $N = 600$) in which the annotated label indicates whether every image depicts a concept of interest. In our setting, these labels were "NSSI" and "no NSSI."

During this part of the training stage, the model observes the full set of images. After the model has observed every image, the performance of the model is computed using an objective function, which considers the output of the model as well as the set of annotated labels for a given image. Then, the parameters of the model are updated based on the model's correct and incorrect predictions. This sequence of steps, usually referred to as a training epoch, is repeated with the goal of gradually improving the performance of the model. In addition, it is common practice to compute, in parallel, the performance on an excluded set of validation images. The latter serves as an indication of the model's performance on unseen pictures. Following the training protocol by Krizhevsky et al. (2012), we sampled "no NSSI" images from the training and validation sets from ILSVRC '12 for our training and testing sets, respectively. Finally, we adopted the classification performance as the core performance metric and applied it to a 48-hr sample of Instagram postings that were automatically scraped in June 2018. We will describe this second computational method in what follows.

Web Scraping of 48 Hr of Instagram Postings Related to NSSI

We used four hashtags as the starting point for the web-scraping procedure. Following Brown et al. (2018), we chose the four NSSI-relevant hashtags #suicide and #cutting, and the German equivalents thereof, which were #selbstmord and #ritzen, respectively. Including German hashtags is arguably more location-specific and they will be less used by international Instagram users, and the inclusion of hashtags related to both cutting and suicidality has been recommended by Brown et al. (2018), since #cutting is often used together with #suicide. On June 25, 2018, we automatically downloaded a complete snapshot of the pictures uploaded to Instagram under these four hashtags over the previous 48 hr using *Python* (Version 3.5.2 for Linux) and the *Selenium* (Selenium Project, 2018; see also Richardson, Berant, & Kuhn, 2018; Grossman, 2017) and *BeautifulSoup* libraries (see Mitchell, 2018).

Web scraping was performed automatically and in reversed chronological order until either a 48-hr limit or a 10,000 most recent posts' limit was reached for each of the four Instagram hashtag feeds (i.e., all posts that were posted on Instagram using one of the four hashtags). For each post, the image file and its time stamp were retained and stored as available through Instagram's public API at that time. In total, $N = 13{,}132$ images that were posted with the NSSI-related hashtags (specifically, #suicide: $n = 7{,}910$, #cutting: $n = 4{,}219$, #selbstmord: $n = 173$, and #ritzen: $n = 830$) were stored. This research protocol was in line with the ethical guidelines as defined by the Association of Internet Research Ethics Working Committee (Markham & Buchanan, 2012).

Results

Accuracy of the Algorithm

During the first training cycle (see top panel in Figure 2), three classes of pictures were predicted. The first two classes were "NSSI-cutting" and "NSSI-no cutting": both were NSSI-related classes, but only "NSSI-cutting" contained images depicting cutting scenes. The third class, "no NSSI," was randomly taken from the validation data of the ILSVRC '12 data set, and this class depicted non-NSSI content. The first training cycle was based on a set of $N = 1{,}384$ images. The second training cycle (bottom panel in Figure 2) only predicted two classes: "NSSI cutting" versus "no NSSI." Like

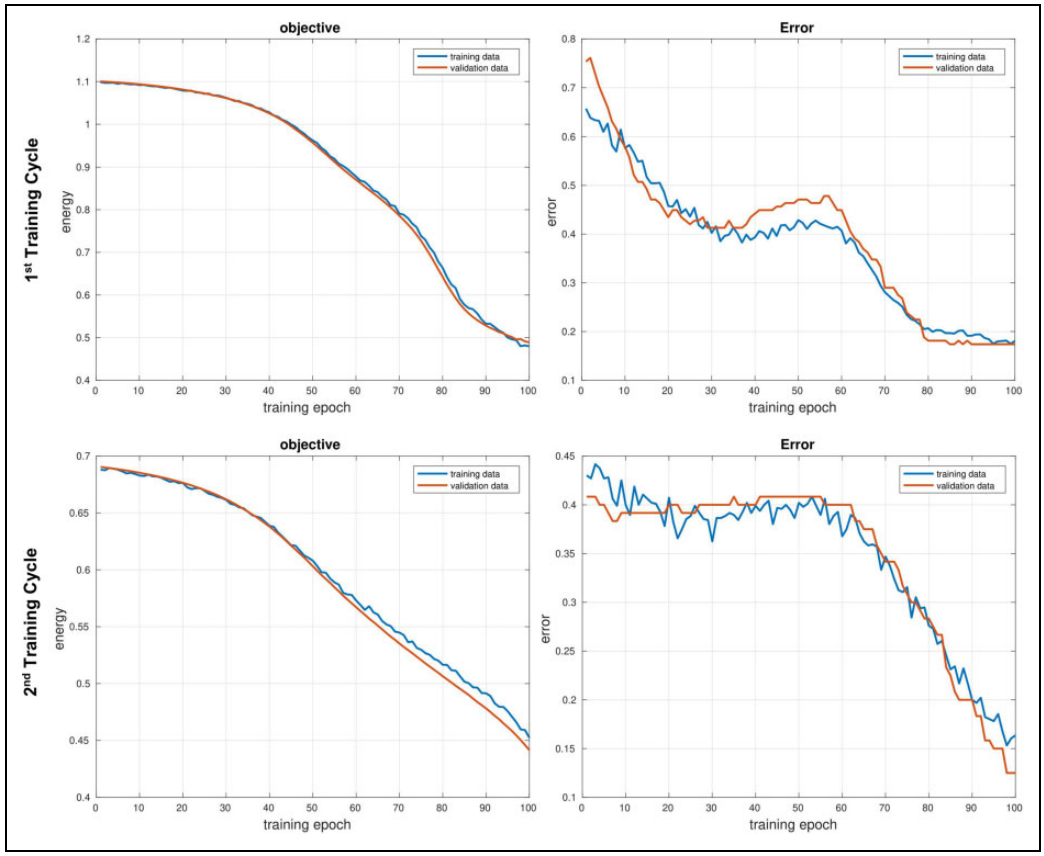

Figure 2. Training/testing results from the first and second cycles of the development of an image-recognition algorithm that automatically detects nonsuicidal self-injury on publicly available Instagram postings for #cutting. On the right, the error rate for the algorithm training (blue line) and the validation (orange line) is shown. On the left, the energy function represents a loss function of the different parameters used in the machine-learning process when developing our algorithm and therefore is indicative of the total error in the NSSI image-recognition task (For interpretation of the references to colours in this figure legend, refer to the online version of this article).

the first training cycle, "no NSSI" images contained content randomly sampled from the ILSVRC '12 data set.

A balanced total set of $N = 1{,}200$ images was used in the second training cycle. As depicted in Figure 2, as the training progressed, the recognition algorithm reached an error rate of $\sim 19\%$ (first training cycle; orange line) and $\sim 13\%$ (second training cycle; orange line), which equals an $\sim 81\%$ and $\sim 87\%$ classification accuracy on the validation data set for the first and second training cycles, respectively. The first training/second training cycles yielded false-positive/false-negative rates of .24/.08 and .23/.07, respectively. Interestingly, after 100 training epochs, the achieved accuracy after the first training cycle nearly converged to a constant. Importantly, the higher performance in the second training cycle was expected, given the lower number of classes that were predicted.

Application in a Large-Scale Real-World Setting

As a second step, we applied the image-recognition algorithm to our real-world setting, represented by the scraped 48-hr sample of NSSI-related pictures posted on Instagram (see Table 1).

Table 1. Language-Related Risk of Exposure to NSSI Pictures on Instagram.

Instagram Hashtag #	NSSI	No NSSI	Language RR	PRE
#cutting	.3003	.6997	1.3883	+38.83%
#ritzen	.4169	.5831		
#suicide	.3989	.6011	0.9997	–0.03%
#selbstmord	.3988	.6012		

Note. Language RR refers to the language-related risk ratio and describes the relative risk of being willingly or unwillingly exposed to NSSI-related images on Instagram depending on the language that is used to search the platform. PRE refers to the percent relative effect of the language that is used to search the platform (interpretation: if #ritzen is searched for on Instagram instead of #cutting, there is a 38.83% higher chance of being willingly or unwillingly exposed to NSSI-related images). The Instagram hashtags #cutting (German: #ritzen) and #suicide (German: #selbstmord) were scraped and automatically analyzed using computational methods. All available pictures posted on Instagram during a 48-hr time window in June 2018 built the basis for these analyses ($N = 13{,}132$). NSSI = nonsuicidal self-injury.

We conducted this additional analysis to test the feasibility of the algorithm in a large-scale real-world setting.

This analysis indicates that users who searched for or followed #ritzen (i.e., in German) had a 1.4 times higher risk of encountering pictures that contained graphic depictions of cutting compared to users who searched for or followed the English hashtag #cutting. This finding corresponds to a percent relative effect of +39%. In contrast, virtually the same language risk ratio (1.0) could be observed for the alternative NSSI-related hashtags in both languages (#suicide or #selbstmord). In these cases, there was no language-induced percent relative effect (–0.03%).

Discussion

We used recent advantages in machine learning to examine the prevalence of pictures that explicitly depict people cutting their wrists as a form of NSSI posted under four different NSSI-related hashtag feeds. It has been argued (Brown et al., 2018) that such pictures may contribute to imitational self-harming behaviors in vulnerable user groups. Thus, a low-cost real-time assessment of NSSI-related posts can contribute to suicide-prevention research and practice by, for example, triggering the display of a help box. The primary contribution of this article is the development and application of an image-recognition algorithm that automatically differentiates cutting as a form of NSSI in pictures posted on the platform. This algorithm can automatically identify cutting-related posts with an accuracy of 87%. All training files and the necessary information to employ, replicate, and further develop the algorithm are available upon request from the authors. In comparison, existing scholarly work that used computational methods to automatically recognize cyberbullying in Instagram posts and that relied on feature combinations (i.e., images and captions) yielded an overall accuracy ranging from 54.56% to 68.55% (Zhong et al., 2016). The presented algorithm therefore sets a first benchmark for automatically detecting NSSI on Instagram. Moreover, the second training cycle in particular could be improved by further training on larger data sets. Such training could help to increase the generalizability of predictions.

In addition, we used web-scraping techniques to access all available Instagram posts under the hashtags #cutting (German: #ritzen) and #suicide (German: #selbstmord) over the course of 48 hr in June 2018. We used this data set to test the feasibility of the developed algorithm. For the first time, this study offers an estimate of the risk ratio and percent relative effect of encountering NSSI on Instagram as they depend on the language under which the four potentially dangerous hashtags can be searched for or followed. Our findings suggest that the relative risk of finding cutting as a form of NSSI on Instagram is 39% higher if the German hashtag (#ritzen) is used compared to the English

hashtag (#cutting). Moreover, NSSI appears to be more often categorized as #suicide than #cutting when the English hashtag is used, while the opposite pattern is true for German-language hashtags (i.e., #ritzen [= cutting] > #selbstmord [= suicide]). Importantly, people choose hashtags themselves, and apparently, for NSSI-related pictures, hashtags along the lines of more (#suicide) or less intentional self-harm (#cutting) are chosen by Instagram users. #suicide (or #selbstmord) could be seen as reflecting a possibly higher conscious suicidal intent, whereas #cutting (or #ritzen) might be more reflective of the nonsuicidal aspects of NSSI. Given the importance of language for NSSI-related stigma, on the one hand, and facilitating NSSI disclosure and help-seeking, on the other hand (see Hasking & Boyes, 2018), the context in which NSSI posts are likely to appear is relevant. The meaning of language for individual suicidality (Arendt, Scherr, Niederkrotenthaler, & Till, 2018) and in algorithm-shaped online environments (Scherr, Haim, & Arendt, 2019) has only recently begun to attract more scholarly attention. The present findings can thus be regarded as a continuation of this line of research, particularly shedding light on the role of language on users' self-categorized NSSI-related behaviors that are publicly shared online. Moreover, therapists could profit from this research by better understanding why some patients use specific language to publicly self-categorize. Knowledge about this appears to still be limited among professionals dealing with NSSI (Hasking & Boyes, 2018). Our approach might spark a healthy discussion in this regard.

The findings may be as they are because the platform's implemented English-language internal warning mechanisms, algorithmic monitoring, and content warnings are more developed than they are for the German-speaking market. Similar findings have recently been reported for the implementation of a suicide-prevention help box on Google that differed significantly across and within countries, depending on the language used to operate the platform services (Scherr et al., 2019). However, the findings related to NSSI on Instagram refer to publicly accessible profile posts. For instance, Arendt (2019) only looked at the German #selbstmord (suicide) hashtag and found through extensive manual coding that within 6 days, almost half of all posts under this hashtag included words or visuals related to suicide. These numbers are largely comparable to the findings of the present study that were based on automated, computational techniques. Importantly, given the growing amount of Finstas (fake Instagram accounts, under which users post less "filtered" content that they are too afraid to post on their "real" accounts, usually shared with fewer people than the real Instagram accounts are), the number of potentially dangerous, NSSI-related pictures might be higher. Often, users also take advantage of the platform by using it to communicate with peers about sensitive topics related to NSSI, which may prevent immediate help. Thus, the automated recognition of potentially harmful pictures on Instagram goes beyond the content warning, which is a platform reaction to certain *hashtag* requests: Helpful NSSI-protective information (such as contact with a crisis intervention center) can be shown for *individual posts*.

Importantly, an image-recognition algorithm seems useful for different groups of Instagram users; namely, those who purposefully search for NSSI images and those who encounter them accidentally while searching or browsing for other content. Potentially beneficial for both groups, the algorithm could be implemented in existing display algorithms and permanently filter content that is knowingly linked with specific harmful hashtags, but it could also be used to optimize the display of the content warning message including links to helplines in an ad hoc manner. Importantly, implications for both scenarios differ. For example, when Instagram users purposefully follow specific content on Instagram by following a specific hashtag, it would be particularly interesting to further investigate the notion of potential echo chambers and the related effects of reinforcing spirals (Slater, 2007, 2015). Such an investigation would be especially relevant for the possible discovery of like-minded NSSI communities on Instagram. For example, 10- to 17-year-olds who visited websites that encourage self-harm were 11 times more likely to think about hurting themselves after controlling for other NSSI risk factors, showing that such content alone is an important risk factor that operates in conjunction with individual predispositions (Mitchell et al.,

2014). Relatedly, an increasing body of research on the functioning of hashtags on Twitter shows that hashtags are powerful in facilitating the expression of nonmainstream political or social sentiments (see Bruns & Burgess, 2015) and bridging individuals with shared interests (Bruns, Moon, Paul, & Münch, 2016, p. 21). Meraz (2017) argues that hashtags facilitate homophilic tendencies and thereby reduce informational diversity and adversarial viewpoints in what then becomes a necessarily consonant virtual environment, while Ging and Garvey (2017) show that Instagram's cross-tagging facility actually opened a pro-ana community to other related hashtags, which might lead to "hashtag dilution" rather than to echo chamber effects. Against the backdrop of the present study, this theorizing would be particularly worrisome if Instagram hashtag filtering unites individuals with a shared interest in NSSI. The potentially reinforcing effects within virtual NSSI communities should not be underestimated given the behavioral consequences of the Werther effect (Niederkrotenthaler et al., 2012; Phillips, 1974; Stack, 2009), but they could just as easily be employed to provide help (see, e.g., Scherr & Reinemann, 2016). Despite being predominantly "weak-tie" support, offers to direct messaging and chats are common on Instagram and might as well be used to reinforce positive attitudinal or behavioral changes to ultimately encourage help seeking, thus being supportive overall of coping with NSSI (see also Jacob, Evans, & Scourfield, 2017). Moreover, it seems worthwhile to further explore to what extent image sharing can be regarded as an expression of gendered distress on Instagram, given the overlap of NSSI-related posts with posts related to eating disorders (Ging & Garvey, 2017) or so-called pain memes (Dobson, 2015).

An automated image-recognition algorithm could not only be used to systematically explore existing hashtags but also to assess trending ones for their potential harmfulness. An ad hoc implementation could, for example, be additionally used within web browsers or picture-based smartphone applications not only to protect vulnerable populations from purposeful exposure to potentially harmful content but also to prevent nonvulnerable users from accidentally accessing unwanted and potentially harmful NSSI content.

The practical implications of the findings presented here are manifold: We make the algorithm available to the interested readers of this article, and we cordially invite other researchers and developers devoted to NSSI and suicide prevention to use, implement, and further develop it. If integrated into Instagram, the algorithm could be used to improve the existing policy for content warnings, which is primarily tied to problematic hashtag requests. Our study thereby advances the area of content protection through awareness messages and moves that content protection from a hashtag-focused function (now) to actual posted-content-focused functioning. This progress is not only relevant to the protection of users from exposure to NSSI-related content but also enhances the overall user experience on the platform. For example, we mentioned #cat in the introduction. There is anecdotal evidence that users post cutting-related content under the hashtag #cat because cats can cause cut-like wounds. Entering this "borrowed" hashtag may lead to children who want to see cute cats being confronted with disturbing cutting pictures. This possibility cannot be adequately addressed with hashtag-focused content warnings. Only a content warning following our posted-content-focused approach can help prevent this situation.

The implementation of our image-recognition algorithm into the display rules of existing content warnings for potentially harmful content would allow the individual filtering of content on Instagram before it is displayed to users (depending on their age, preferences, or past problematic use patterns). Hence, this algorithm could contribute to a more supportive online environment on a very popular picture-based social networking platform.

Limitations

The application presented in this study focused on images of cutting on Instagram. The algorithm needs to be trained further using a larger and representative data sample that covers all possible

specific patterns on pictures that reflect, for example, different common schemes in NSSI postings. For instance, Brown et al. (2018) reported that users posted NSSI pictures that showed whole words cut into those users' extremities, and Arendt (2019) discussed preliminary observations of fast cutting techniques in audiovisual material. In general, the algorithm presented here can handle complex elements and video material in addition to images. However, this process requires further training and the conversion of video material into stills that are then algorithmically analyzed. Furthermore, algorithm training using specific subsets of NSSI content on Instagram is needed, as are alternative decoding and conversion programming techniques. However, the present study's first benchmark for automatically identifying NSSI on Instagram could and should be further improved. Readers should also bear in mind that the current data collection procedure was conducted in reverse chronological order over the 48-hr data collection period. Thus, the built-in Instagram algorithm for detecting and autodeleting harmful content may have already eliminated some of the NSSI content. Moreover, it is possible that the NSSI-related content was spiking at the time of the data collection. Instagram hashtags tend to have a specific life cycle (see Ging & Garvey, 2017) where they become more ("going viral") and then less popular ("becoming mainstream"), with new hashtags being used or adapted, which cannot be ruled out here. Nevertheless, there still seems to be enough room for improvement, given the current sample's prevalence of NSSI in Instagram images.

Conclusion

Research has only begun to look closely at potentially dangerous NSSI-related Instagram hashtags that were heavily restricted via content accessibility and manual coding of the predominantly visual material. By combining web-scraping techniques to make large amounts of Instagram content automatically available and analyzing those contents using a newly developed automatic image-recognition algorithm, we aimed to give new impetus to this emerging line of research (e.g., Arendt, 2019; Brown et al., 2018). Further, developing the introduced algorithm is strongly suggested and supported through the authors making all the materials from this study available online.

Authors' Note

We would like to thank Jian Yang for having inspired us to conduct this research. All data related to this study are available upon request from the authors. In this study, the software *4K Stogram* (https://www.4kdownload.com/products/product-stogram), *Selenium* (Selenium Project, 2018; https://www.seleniumhq.org/), and *BeautifulSoup* (Richardson, 2018; https://www.4kdownload.com/products/product-stogram) were used to obtain images from Instagram. The image-recognition algorithm was developed using the *Maconvnet* framework (Vedaldi & Lenc, 2015; http://www.vlfeat.org/matconvnet/). The *AlexNet architecture* was used as the neural network classifier for the image-recognition algorithm.

Declaration of Conflicting Interests

The author(s) declare no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, *50*, 179–211. doi:10.1016/0749-5978(91)90020-T

Ajzen, I., & Fishbein, M. (1973). Attitudinal and normative variables as predictors of specific behavior. *Journal of Personality and Social Psychology*, *27*, 41–57. doi:10.1037/h0034440

Arendt, F. (2019). Suicide on Instagram: Content analysis of a German suicide-related hashtag. *Crisis: The Journal of Crisis Intervention and Suicide Prevention*, *40*, 36–41. doi:10.1027/0227-5910/a000529

Arendt, F., Scherr, S., Niederkrotenthaler, T., & Till, B. (2018). The role of language in suicide reporting: Investigating the influence of problematic suicide referents. *Social Science & Medicine*, *208*, 165–171. doi: 10.1016/j.socscimed.2018.02.008

Baker, T. G., & Lewis, S. P. (2013). Responses to online photographs of non-suicidal self-injury: A thematic analysis. *Archives of Suicide Research*, *17*, 223–235. doi:10.1080/13811118.2013.805642

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*, 337–346. doi:10.1017/S0033291717001751

Bruns, A., & Burgess, J. (2015). Twitter hashtags from ad hoc to calculated publics. In N. Rambukkana (Ed.), *Hashtag publics: The power and politics of discursive networks* (pp. 13–28). New York, NY: Lang.

Bruns, A., Moon, B., Paul, A., & Münch, F. (2016). Towards a typology of hashtag publics: A large-scale comparative study of user engagement across trending topics. *Communication Research and Practice*, *2*, 20–46. doi:10.1080/22041451.2016.1155328

De Riggi, M. E., Moumne, S., Heath, N. L., & Lewis, S. P. (2016). Non-suicidal self-injury in our schools: A review and research-informed guidelines for school mental health professionals. *Canadian Journal of School Psychology*, *32*, 122–143. doi:10.1177/0829573516645563

Dobson, A. S. (2015). Girls' "pain memes" on YouTube: The production of pain and femininity in a digital network. In S. Baker, B. Robards, & B. Buttigieg (Eds.), *Youth cultures and subcultures: Australian perspectives* (pp. 173–182). Farnham, England: Ashgate.

Dyson, M. P., Hartling, L., Shulhan, J., Chisholm, A., Milne, A., Sundar, P., … Newton, A. S. (2016). A systematic review of social media use to discuss and view deliberate self-harm acts. *PLos One*, *11*, e0155813. doi:10.1371/journal.pone.0155813

Fahey, R. A., Matsubayashi, T., & Ueda, M. (2018). Tracking the Werther effect on social media: Emotional responses to prominent suicide deaths on twitter and subsequent increases in suicide. *Social Science & Medicine*, *219*, 19–29. doi:10.1016/j.socscimed.2018.10.004

Ging, D., & Garvey, S. (2017). "Written in these scars are the stories I can't explain": A content analysis of pro-ana and thinspiration image sharing on Instagram. *New Media & Society*, *20*, 1181–1200. doi:10.1177/1461444816687288

Grossman, T. (2017). My open source Instagram bot got me 2,500 real followers for $5 in server costs. Retrieved from https://medium.freecodecamp.org/my-open-source-instagram-bot-got-me-2-500-real-followers-for-5-in-server-costs-e40491358340

Hamza, C. A., Stewart, S. L., & Willoughby, T. (2012). Examining the link between nonsuicidal self-injury and suicidal behavior: A review of the literature and an integrated model. *Clinical Psychology Review*, *32*, 482–495. doi:10.1016/j.cpr.2012.05.003

Hasking, P., & Boyes, M. (2018). Cutting words: A commentary on language and stigma in the context of nonsuicidal self-injury. *The Journal of Nervous and Mental Disease*, *206*, 829–833. doi:10.1097/NMD.0000000000000899

Instagram. (2017). Now you can follow hashtags on Instagram. Retrieved from https://instagram-press.com/blog/2017/12/12/now-you-can-follow-hashtags-on-instagram/

Jacob, N., Evans, R., & Scourfield, J. (2017). The influence of online images on self-harm: A qualitative study of young people aged 16–24. *Journal of Adolescence*, *60*, 140–147. doi:10.1016/j.adolescence.2017.08.001

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems* (Vol. 25, pp. 1097–1105). Retrieved from https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

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*, 249–262. doi:10.1002/jclp.22242

Markham, A., & Buchanan, E. (2012). Ethical decision-making and Internet research: Recommendations from the AoIR ethics working committee (Version 2.0). Retrieved from https://aoir.org/reports/ethics2.pdf

Mars, B., Heron, J., Crane, C., Hawton, K., Kidger, J., Lewis, G., . . . Gunnell, D. (2014). Differences in risk factors for self-harm with and without suicidal intent: Findings from the ALSPAC cohort. *Journal of Affective Disorders*, *168*, 407–414. doi:10.1016/j.jad.2014.07.009

Meraz, S. (2017). Hashtag wars and networked framing: The private/public networked protest repertoires of occupy on twitter. In A. S. Tellería (Ed.), *Between the public and private in mobile communication* (pp. 303–323). New York, NY: Routledge.

Miguel, E. M., Chou, T., Golik, A., Cornacchio, D., Sanchez, A. L., DeSerisy, M., & Comer, J. S. (2017). Examining the scope and patterns of deliberate self-injurious cutting content in popular social media. *Depression and Anxiety*, *34*, 786–793. doi:10.1002/da.22668

Minkkinen, J., Oksanen, A., Kaakinen, M., Keipi, T., & Räsänen, P. (2017). Victimization and exposure to pro-self-harm and pro-suicide websites: A cross-national study. *Suicide and Life-Threatening Behavior*, *47*, 14–26. doi:10.1111/sltb.12258

Mitchell, R. (2018). *Web scraping with python: Collecting more data from the modern web cover* (2nd ed.). Sebastopol, CA: O'Reilly Media.

Mitchell, K. J., Wells, M., Priebe, G., & Ybarra, M. L. (2014). Exposure to websites that encourage self-harm and suicide: Prevalence rates and association with actual thoughts of self-harm and thoughts of suicide in the United States. *Journal of Adolescence*, *37*, 1335–1344. doi:10.1016/j.adolescence.2014.09.011

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

Niederkrotenthaler, T., Fu, K. W., Yip, P. S. F., Fong, D. Y. T., Stack, S., Cheng, Q., & Pirkis, J. E. (2012). Changes in suicide rates following media reports on celebrity suicide: A meta-analysis. *Journal of Epidemiology and Community Health*, *66*, 1037–1042. doi:10.1136/jech-2011-200707

Nock, M. K., & Favazza, A. (2009). Non-suicidal self-injury: Definition and classification. In M. K. Nock (Ed.), *Understanding non-suicidal self-injury: Origins, assessment, and treatment* (pp. 9–18). Washington, DC: American Psychological Association.

Nock, M. K., & Prinstein, M. J. (2004). A functional approach to the assessment of self-mutilative behavior. *Journal of Consulting and Clinical Psychology*, *72*, 885–890. doi:10.1037/0022-006x.72.5.885

O'Connor, R. C., Rasmussen, S., & Hawton, K. (2009). Predicting deliberate self-harm in adolescents: A six month prospective study. *Suicide and Life-Threatening Behavior*, *39*, 364–375. doi:10.1521/suli.2009.39.4.364

Phillips, D. P. (1974). The influence of suggestion on suicide: Substantive and theoretical implications of the Werther effect. *American Sociological Review*, *39*, 340–354.

Radovic, S., & Hasking, P. (2013). The relationship between portrayals of nonsuicidal self-injury, attitudes, knowledge, and behavior. *Crisis: The Journal of Crisis Intervention and Suicide Prevention*, *34*, 324–334. doi:10.1027/0227-5910/a000199

Richardson, K., Berant, J., & Kuhn, J.(2018). Polyglot semantic parsing in APIs. Paper presented at the NAACL-HLT 2018, New Orleans, LO. Retrieved from https://arxiv.org/pdf/1803.06966.pdf

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., . . . Fei-Fei, L. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*, 211–252.

Scherr, S., & Reinemann, C. (2016). First do no harm: Cross-sectional and longitudinal evidence for the impact of individual suicidality on the use of online health forums and support groups. *Computers in Human Behavior*, *61*, 80–88. doi:10.1016/j.chb.2016.03.009

Scherr, S., Haim, M., & Arendt, F. (2019). Equal access to online information? Google's suicide-prevention disparities may amplify a global digital divide. *New Media and Society*, *21*, 562–582. doi:10.1177/1461444818801010

Selenium Project. (2018). SeleniumHQ Browser Automation. Retrieved from https://www.seleniumhq.org/

Slater, M. D. (2007). Reinforcing spirals: The mutual influence of media selectivity and media effects and their impact on individual behavior and social identity. *Communication Theory*, *17*, 281–303. doi:10.1111/j.1468-2885.2007.00296.x

Slater, M. D. (2015). Reinforcing spirals model: Conceptualizing the relationship between media content exposure and the development and maintenance of attitudes. *Media Psychology*, *18*, 370–395. doi:10.1080/15213269.2014.897236

Smith, A., & Anderson, M. (2018). Social media use in 2018: A majority of Americans use Facebook and YouTube, but young adults are especially heavy users of Snapchat and Instagram. Retrieved from http://www.pewinternet.org/2018/03/01/social-media-use-in-2018/

Stack, S. (2009). Copycat effects of fictional suicide: A meta-analysis. In S. Stack & D. Lester (Eds.), *Suicide and the creative arts* (pp. 231–243). New York, NY: Nova.

Vedaldi, A., & Lenc, K. (2015). *MatConvNet: Convolutional neural networks for MATLAB*. Paper presented at the 23rd ACM International Conference on Multimedia (MM '15), New York, NY.

Wang, X., Wei, F., Liu, X., Zhou, M., & Zhang, M. (2011). *Topic sentiment analysis in twitter: A graph-based hashtag sentiment classification approach*. Paper presented at the Proceedings of the 20th ACM international conference on Information and knowledge management, Glasgow, Scotland, UK.

Zhong, H., Li, H., Squicciarini, A., Rajtmajer, S., Griffin, C., Miller, D., & Caragea, C. (2016). *Content-driven detection of cyberbullying on the Instagram social network*. Paper presented at the Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI-16).

Author Biographies

Sebastian Scherr, PhD, is an assistant professor at the School for Mass Communication Research, University of Leuven, Belgium. His research interests focus on differential media uses and effects in health communication and political communication, with a special emphasis on suicide prevention and empirical methods. Email: sebastian.scherr@kuleuven.be

Florian Arendt, PhD, holds the Tenure Track Professorship in Health Communication at the Department of Communication, University of Vienna, Austria. His research focuses on health communication with a special emphasis on suicide prevention. Email: florian.arendt@univie.ac.at

Thomas Frissen is a PhD candidate at the Institute for Media Studies in KU Leuven, Belgium. His areas of interest are online networks as facilitators for extremist ideologies and the development of new methods to study them. Email: thomas.frissen@kuleuven.be

José Oramas M., PhD, is a postdoctoral researcher at the Department of Electrical Engineering (ESAT-PSI), University of Leuven. His research focuses on the automatic interpretation/explanation of models addressing artificial visual perception tasks. Email: jose.oramas@kuleuven.be