End violence:
Women's rights and safety online

Internet intermediaries and violence against women online

Twitter: A case study

Carly Nyst

Association for Progressive Communications (APC)

July 2014

Ministry of Foreign Affairs

This research is part of the APC "End violence: Women´s rights and online safety" project funded by the Dutch Ministry of Foreign Affairs (DGIS) and is based on a strong alliance with partners in seven countries: Bosnia and Herzegovina, Colombia, Democratic Republic of Congo, Kenya, Mexico, Pakistan and the Philippines. For more information visit GenderIT.org and Take Back the Tech!

Table of contents

Introduction to the research

This profile is part of a short study of the policies of three major internet intermediaries, Facebook, YouTube and Twitter, with respect to violence against women online. The study aims to map the corporate policies of these intermediaries that allow for the identification, reporting and rectification of incidents of harassment or violence against women via the service that the intermediary provides. In addition to providing a detailed summary of the user policies relevant to this issue, the study also compares the impact and effectiveness of those policies against the framework of the UN Guiding Principles on Business and Human Rights. The study was conducted on the basis of desk research and an analysis of corporate policies and terms of service, and interviews with representatives of the intermediaries. However, YouTube was the only company out of the three to agree to an interview with the researchers.

About Twitter

Twitter is an online social networking and micro-blogging service, established in 2006, with more than 200 million users and 3,000 employees. It is headquartered in California, and brought in upwards of USD 664 million in 2013. Outside of the US, Twitter has offices in Amsterdam, Berlin, Dublin, London, Madrid, Paris, Rio de Janeiro, São Paulo, Singapore, Sydney, Seoul, Tokyo, Toronto and Vancouver. The countries with the highest percentage of Twitter users per capita are Saudia Arabia, Indonesia, Spain, Venezuela and Argentina; the UK, US and Netherlands are also in the top 10.[1]

Report format

This report is broken down into three main sections:

1. Critical analysis of main trends.
2. Charting the impact and effectiveness of Twitter's policies and procedures with respect to violence against women, using the framework of the Guiding Principles on Business and Human Rights.
3. A detailed breakdown of Twitter's user policies, redress mechanisms and the evolution of its approach to violence against women.

[1] blog.peerreach.com/2013/11/4-ways-how-twitter-can-keep-growing

1. Analysis of main trends

Twitter has had revolutionary impacts upon the ability for expression, ideas and political movements to spread and connect across the internet. Twitter acts as a tool for communication, education and expression for more than 200 million users, who send more than 400 million Tweets each day.[2] It has played a number of key social roles, from facilitating protest organisation during revolutions in Egypt, Tunisia and Iran, to enabling humanitarian responses and analysis for crises in Haiti and the Philippines.

Twitter has taken a consistent stand on the right to freedom of expression and opinion, often to the detriment of steps to counter abuse and hate speech. It has resisted cooperating with law enforcement authorities in response to hate speech allegations regarding the use of Twitter by French users to propagate anti-Semitic speech in contravention of French law, a battle which Twitter ultimately lost in the French courts.[3] Their user policies place the onus on the user to report to law enforcement and make no strong statements about the unacceptability of hate speech or violence against women. The only reportable violations under Twitter's policies are those relating to abuse and threats of violence, which are not well defined. Anecdotal evidence that emerged in summer 2013, when numerous high-profile women spoke out about the violent threats they were receiving via Twitter, shows that Twitter's reporting processes have not been working effectively to prevent and redress violence against women via the platform.

There is little information available about how Twitter's reporting processes work, and the company does not publish disaggregated statistics about the types of reports it receives. Moreover, Twitter does not publicly engage with civil society about its reporting process or seek input from affected groups. The negative implications of failing to take a consultative approach were evidenced most clearly in 2013 when Twitter reversed changes to the "blocking" mechanism[4] that it had announced only the previous day. The changes had allowed blocked users to continue viewing the tweets of and interact with accounts that had blocked them, while remaining invisible so that the reporting user could not see that the blocked user was following them. Twitter said that the changes were made so that the blocked user would not know they had been blocked, a development that often leads blocked users to retaliate against the reporting user. However, within hours of announcing the changes, Twitter was flooded with the complaints of angry users, and confronted with an online petition to reverse the change.[5] Twitter reversed its changes almost immediately, reverting to the previous policy. It is clear in such situations that engagement with public policy, public ethics and laws is a necessary means of ensuring that the platform responds to the needs of its users.

[2]Moore, H. (2013, September 12). Twitter heads for stock market debut by filing for IPO. *The Guardian*. www.theguardian.com/technology/2013/sep/12/twitter-ipo-stock-market-launch?CMP=EMCNEWEML6619I2&et_cid=48826&et_rid=7107573&Linkid=http%3a%2f%2fwww.theguardian.com%2ftechnology%2f2013%2fsep%2f12%2ftwitter-ipo-stock-market-launch

[3]Beardsley, E. (2013, January 22). French Twitter lawsuit pits free speech against hate speech. *NPR*. www.npr.org/2013/01/22/169998834/french-twitter-lawsuit-pits-free-speech-against-hate-speech

[4]BBC. (2013, December 13). Twitter backtracks on blocking changes. *BBC News*. www.bbc.com/news/technology-25361255

[5]Shih, J. (2013, December 13). Twitter users revolt over changes to abusive behavior policy. *Reuters*. www.reuters.com/article/2013/12/13/twitter-abuse-idUSL1N0JS04920131213

2. Compliance with the Guiding Principles on Business and Human Rights

Human rights obligations do not only relate to the actions or omissions of states. Companies are also required under international law to respect human rights, to avoid infringing human rights, and to address adverse human rights impacts with which they are involved. This means not only do they have to take action when they play a role in human rights violations, but they also have to take positive steps to prevent, mitigate and remedy human rights violations.

The steps required by companies to fulfil human rights obligations were analysed at length by the UN Special Representative of the Secretary-General on the issue of human rights and transnational corporations and other business enterprises. The SRSG compiled a set of Guiding Principles on Business and Human Rights, which was endorsed by the UN Human Rights Council in 2011. The Guiding Principles enshrine a framework of obligations, entitled "Protect, Respect and Remedy", which tells both states and companies what steps they should take to promote human rights.

When it comes to addressing technology-related violence against women, the second pillar of the Framework – Respect – provides guidance for internet intermediaries as to the actions they should take to ensure that women's rights online are promoted and respected. The Respect pillar sets a number of benchmarks that companies must reach in order to be in compliance with human rights obligations.

The third pillar of the Framework – Access to Remedy – establishes that states must take steps to ensure judicial, administrative or other remedies to ensure that victims of human rights abuses can obtain redress. While the pillar primarily addresses the role of states in this regard, it also provides that business enterprises should establish or participate in effective operational-level grievance mechanisms for adverse human rights abuses.

The Guiding Principles may provide an effective and useful structure within which to engage internet intermediaries on technology-related violence against women concerns. The Framework prescribes a number of actions that can be adapted to analyse the actions of internet intermediaries in this regard.

We have developed a list of questions,[6] to correspond with the Principles, that organisations, advocates and activists can use to analyse the compliance of internet intermediaries with the Guiding Principles. Below, we use the questions to address the question of Twitter's compliance with human rights obligations in its approach to issues of violence against women.

[6] These questions are based on the Access to Justice Framework for Corporate Policies, as detailed in the Ending Violence: Domestic Legal Remedies and Corporate Policies/Redress Mechanisms Research Design.

Policy commitment	
1.	Does the intermediary have a publicly available statement of policy that stipulates the organisation's policy with respect to violence against women (in all of its forms)?
	Twitter does not publicly assert its position with respect to violence against women in any of its user policies or guides. It does not mention hate speech in any of its policies. In response to the summer 2013 allegations regarding violence against women on Twitter, Twitter representatives have made public statements to the effect that such behaviours are not acceptable and that Twitter can be doing more to counter them.[7] Beyond that, however, Twitter has done little to provide information about hate speech and violence against women, publicise a strong stance against such behaviour, or provide clear and concrete avenues of redress. Twitter should amend its Twitter Rules to emphasise that it does not permit hate speech, graphic or gratuitous violence, threats, predatory behaviour, harassment or the invasion of privacy, and to make a public commitment to the promotion of human rights standards beyond just the encouragement of free speech. Twitter should make available policies that explicitly address gender-related violence or harassment and take a strong stance on respect for diversity and for women's rights.
2.	Has the intermediary taken due diligence steps to understand the way in which it may be facilitating violence against women, in order to inform its policies and procedures?
	Prior to summer 2013, there was no indication that Twitter sought or took into account the input of stakeholders or community groups, or commissioned studies or due diligence, on issues related to harassment of and violence against women. Twitter does not appear to have a stakeholder input process for any of its policies, nor to seek consultation on any of its technical or policy changes as they occur. In response to the summer 2013 outrage against Twitter's inaction in the face of violence against women, the intermediary expressed interest in receiving the input of stakeholder groups and opening a dialogue with the women's rights communities; however, it is unclear whether these initiatives have been pursued. Twitter should take a more proactive stance to the issue of violence against women via its platform. It is insufficient for it to address these issues only when a scandal flares in the media. Twitter should undertake a comprehensive consultative investigation of the ways in which it might facilitate and address violence against women online.
Due diligence	
3.	Has the intermediary engaged in meaningful consultation with women, either by soliciting the input of users or by engaging women's rights groups and activists, to understand the potential adverse impacts of its services on women's rights?
	Twitter does not appear to have in place a process of consultation with women's rights groups or activists, nor does it point to any concrete examples of when it has reached out to such groups for input. Twitter should be strongly recommended to improve its processes in this respect.
4.	Is responsibility for addressing issues of violence against women assigned to the appropriate level and function within the intermediary?
	Twitter does not seem to have mainstreamed women's issues into its policies or procedures; there is no particular person or division responsible for women's issues and no working group, stream or committee engaged in these issues. Twitter has not published a policy on or approach to violence against women issues, nor has it signed the Women's Empowerment Principles. With respect to Twitter's reporting processes, it is unclear whether issues of violence

[7] Ensor, J. (2013, August 3). Twitter boss personally apologises to female victims of abuse. *The Telegraph*. www.telegraph.co.uk/technology/twitter/10220410/Twitter-boss-personally-apologises-to-female-victims-of-abuse.html

	against women are given any serious consideration at all, as there is no clear explanation available as to how such concerns would be dealt with; the only information we were able to uncover was an interview given by a Twitter representative. Twitter should urgently amend its policies to include explicit reference to violence against women or hate speech on the basis of gender. It should publish detailed information about how violence against women-related complaints are handled and what standards are applied to them, as well as disaggregated information about the gender, expertise and training of complaints handlers.
5.	Do internal decision-making processes enable effective responses to issues of violence against women?
	Anecdotal evidence that emerged as a result of the summer 2013 accusations shows that there is no effective internal response mechanism in Twitter to ensure that issues of violence against women are addressed efficiently, or at all. This may have improved since that time, given that awareness about the use of Twitter to propagate violence and harassment against women has no doubt increased, but there is no further evidence available in that regard. Twitter must embrace greater transparency of its reporting processes in order to enable greater scrutiny of them.
6.	Does the intermediary track how effective its responses to issues of violence against women are, either by tracking indicators or seeking feedback from affected stakeholders?
	Twitter takes a hands-off approach to the content that its users disseminate via its platform. This attitude is clear throughout its policies, which throughout stipulate that the user should report serious concerns to law enforcement and that Twitter will not actively monitor its platform to seek out abuse. It does not provide any information about the amount of reports it receives, how many accounts it blocks and how many incidents are successfully resolved. Without this information it is impossible to assess how effective its responses might be.
7.	Does the intermediary publicly communicate both the occurrence of, and its response to, issues of violence against women?
	Twitter does not monitor or publish information about instances of violence against women on its platform or about the actions it takes to mitigate or address instances of violence against women. This is a serious failing which should be rectified by Twitter.
Remediation	
8.	Is there a grievance mechanism in place for individuals or communities who are adversely affected by violence against women?
	Twitter adopts a self-reporting model, where users are invited to complete an online form if: a) someone is posting private information about them; b) someone on Twitter is being abusive; or c) someone on Twitter is sending violent threats. Submitting this form requires users to provide significant information about the date, time and URLs of the tweets in question, as well as any other background information. The user is required to provide their email address and their Twitter handle, and is warned of the consequences of making false or vexatious complaints. Twitter also allows users to block other users, and to report media or individual tweets directly from their source by using a reporting button that was introduced in mid-2013. This avenue is only available on certain hardware and platforms – initially it was only available on iOS, but in late 2013 Twitter began rolling it out for Android, and on the web.
9.	Does the intermediary consult stakeholder groups on the design and performance of its grievance mechanism?
	There is no indication that Twitter consulted women's groups or activists in the design of its grievance mechanism, nor is there any provision for feedback from the public and civil society on how the mechanism could be better designed or operated.

10.	Does the mechanism meet the following effectiveness criteria?
	10a. Legitimacy – is the mechanism viewed as trustworthy, and is it accountable to those who use it?
	The only publicly available information from those who have used the feedback mechanism emerged during the summer 2013 incidents. Feminist media critic Anita Sarkeesian tweeted, "I've reported numerous rape threats to @Twitter. This is how they respond: 'The account is currently not in violation of the Twitter Rules.'" She also posted a screen grab of one such tweet, showing the account @CoolDehLa tweeting, on 26 December 2012, "@femfreq I will rape you when i get the chance." Sarkeesian wrote, "Twitter says 'We have found the reported account is currently not in violation of the Twitter Rules at this time.'" Sarkeesian's two tweets were retweeted more than 7,000 times.[8] In response to Sarkeesian's and other's public criticism of Twitter, the intermediary has introduced the reporting button that allows for users to more easily and simply report tweets at their source. There is insufficient anecdotal or other evidence to establish whether these changes have improved the effectiveness of Twitter's handling of reports. Twitter should thus publish its own information on the number of reports received, responded to, and acted upon, in order to enable more detailed engagement with this question.
	10b. Accessibility – is the mechanism easily located, used and understood?
	Since July 2013, the Twitter reporting mechanism has been easily locatable, and can be accessed underneath each tweet. Users are requested to select from Spam, Compromised and Abusive, and if they choose Abusive they are directed to a more rigorous form. This form can also be accessed through the Twitter Support Center. The introduction of the reporting button is extremely important, as accessing the form via the Twitter Support Center can be difficult – while it is easily locatable via Google it is not immediately visible on the service, nor through any of the applications or platforms that users might access Twitter through (like TweetDeck), and even once a user locates the Support Center, there is significant reading and searching required to find the location of the I'm reporting an abusive user page. The Twitter reporting mechanism is easily understood and is written in plain language and users are guided through the steps quickly and simply. As of mid-2012, Twitter supported 30 languages, providing policies in as many languages.[9] There might be a slight difficulty for users in providing the URLs of certain tweets, but Twitter provides assistance in this regard. It appears that individuals are only able to report violations online, and this may exclude individuals without access to computers or without computer literacy from accessing the grievance mechanism. While law enforcement and government agencies may contact Twitter offline through their legal department, individuals are directed to use "regular support methods" of online forms. Furthermore, Twitter's complaint process requires the disclosure of at least basic identifying information, and seems to require that the reporter has a Twitter account. There is no strong statement about whether Twitter will disclose identifying information to the reportee; accordingly, this requirement may deter women from using the complaints procedure.

[8]Greenhouse, E. (2013, August 1). Twitter's free-speech problem. *The New Yorker*. www.newyorker.com/online/blogs/elements/2013/08/how-free-should-speech-be-on-twitter.html

[9]Long, M. C. (2012, July 5). Twitter now supports 30 different languages. *AllTwitter*. www.mediabistro.com/alltwitter/30-different-languages-twitter_b25046

	10c. Predictability – is there a clear and open procedure with indicative time frames, clarity of process and means of monitoring implementation?
	Twitter gives no public information about how reporting processes work, what time frames are applicable, who will be monitoring complaints, what training they have, what specific standards they will apply, and how a user will be notified of the outcome. There are no examples or case studies available for people to understand how previous complaints have been dealt with. Twitter should introduce greater transparency in this regard. There is significant information about the circumstances under which Twitter will cooperate with law enforcement, which is commendable.
	10d. Equitable – does the intermediary provide sufficient information and advice to enable individuals to engage with the mechanism on a fair and informed basis?
	Twitter provides no extended guidance about what constitutes abusive behaviour or threats of violence for the purpose of its complaint mechanism. This not only undermines efforts to prevent such behaviour, but it prevents users from making their complaint in a way that complies with Twitter's requirements.
	10e. Transparent – are individuals kept informed about the progress of their matter?
	It is unclear to what extent Twitter keeps reporters updated about the progress of their complaints. Twitter should be encouraged to improve these processes, as certainty as to the outcome is an important element of assisting victims/survivors of violence against women.
	10f. Rights-compatible – do the outcomes and remedies accord with internationally recognised human rights?
	There does not appear to be an established process for appealing an adverse decision about complaints; this is supported by the anecdotal accounts provided by Sarkeesian. This would not comport with the need for independence, impartiality and accountability in processes of remedy and redress. There is a serious need for more explicit recognition of issues of violence against women on Twitter and more concrete means of addressing them. There is also need for greater participation of women's groups and activists in the design and implementation of the reporting mechanism. Currently, Twitter is completely shut off to women who have serious concerns about the ways in which violence against women is being facilitated by the intermediary. There are insufficient victim support mechanisms available through Twitter, and this should be rectified. Twitter should put in place better procedures to ensure that individuals and communities can give input into the design of the processes themselves, which do not reflect a consideration of the particular difficulties of ensuring justice for victims/survivors of violence against women online.
	10g. Source of continuous learning – does the intermediary draw on experiences to identify improvements for the mechanism and to prevent future grievances?
	As reiterated above, Twitter has taken few visible measures to incorporate the particular challenges of violence against women into its complaints mechanism, and without greater transparency around the process and greater introspection on this particular point it is difficult to see how it can begin to take proactive steps to prevent further violence against women on the platform.

2.1. Recommendations

1. In order to ensure that it is meeting its obligations to respect and advance human rights standards, particularly the right of women to be free from harassment, hatred and violence online, Twitter should take the following steps:
2. Amend its Twitter Rules to emphasise that it does not permit hate speech, graphic or gratuitous violence, threats, predatory behaviour, harassment or the invasion of privacy, and make a public commitment to the promotion of human rights standards beyond just the encouragement of free speech.
3. Sign the Women's Empowerment Principles.[10]
4. Make available policies that explicitly address gender-related violence or harassment and take a strong stance on respect for diversity and for women's rights.
5. Take a more proactive stance to the issue of violence against women via its platform. It is insufficient for it to address these issues only when a scandal flares in the media. Twitter should undertake a comprehensive consultative investigation of the ways in which it might facilitate and address violence against women online.
6. Put in place a concrete process of consultation with women's rights groups and activists, particularly those outside of the US and Europe, on the design, implementation and evaluation of policies and procedures.
7. Establish a point-person responsible for understanding and responding to issues related to violence against women, and for establishing – in consultation with the relevant individuals and communities – a Twitter policy towards issues of violence against women.
8. Provide regular training to staff responsible for moderation on issues related to human rights in general, and to the specific realities of women's rights as they pertain to health, sexuality, violence.
9. Publish disaggregated information about the gender, expertise and training of complaints handlers dealing with content- and privacy-related complaints.
10. Provide greater transparency about complaints processes more generally, what standards are applied and how complaints are dealt with throughout their lifecycle.
11. Publish information about instances of violence against women on Twitter, including information on the number of reports received, responded to, and acted upon, in order to enable more detailed engagement on these issues.
12. Consider providing alternative reporting mechanisms for individuals who are not computer literate or who might be unwilling to disclose identifying information in making a complaint.

[10] The Women's Empowerment Principles (WEP) is a global initiative focused specifically on corporate social responsibility and women's human rights, and is a collaboration between UN Women and the UN Global Compact (see weprinciples.org). Signing in support of the WEP entails recognition of the costs of violence against women to businesses, and a commitment to developing internal and external initiatives to increase women's empowerment within the workplace, marketplace and community.

13. Make victim support mechanisms available through Twitter.
14. Provide greater information about the time frame within which complaints are dealt with, and the avenues for review of decisions.

3. Twitter's user policies

Below, we analyse Twitter's current user policies to ascertain to what extent they address, prohibit, and provide redress for technology-related violence against women.

3.1. Which rights violations are explicitly recognised and provided for in corporate policies?

Twitter's Terms of Service stipulate users' rights and obligations with respect to their use of the platform.

Under Section 5 of the Terms of Service, Twitter places the responsibility for content posted on the site firmly with the platform's users, stating:

> "You are responsible for your use of the Services, for any Content you provide, and for any consequences thereof, including the use of your Content by other users and our third party partners."

Section 4 of the Terms of Service eschews any liability on the part of Twitter for content posted on the platform:

> 4. Content on the Services
>
> All Content, whether publicly posted or privately transmitted, is the sole responsibility of the person who originated such Content. We may not monitor or control the Content posted via the Services and, we cannot take responsibility for such Content. Any use or reliance on any Content or materials posted via the Services or obtained by you through the Services is at your own risk.
>
> We do not endorse, support, represent or guarantee the completeness, truthfulness, accuracy, or reliability of any Content or communications posted via the Services or endorse any opinions expressed via the Services. You understand that by using the Services, you may be exposed to Content that might be offensive, harmful, inaccurate or otherwise inappropriate, or in some cases, postings that have been mislabeled or are otherwise deceptive. Under no circumstances will Twitter be liable in any way for any Content, including, but not limited to, any errors or omissions in any Content, or any loss or damage of any kind incurred as a result of the use of any Content posted,

emailed, transmitted or otherwise made available via the Services or broadcast elsewhere.

Section 8 introduces the Twitter Rules, which govern limitations on the type of content that can be used on the service. The section notes:

> 8. Restrictions on Content and Use of the Services
>
> Please review the Twitter Rules (which are part of these Terms) to better understand what is prohibited on the Service. We reserve the right at all times (but will not have an obligation) to remove or refuse to distribute any Content on the Services, to suspend or terminate users, and to reclaim usernames without liability to you. We also reserve the right to access, read, preserve, and disclose any information as we reasonably believe is necessary to (i) satisfy any applicable law, regulation, legal process or governmental request, (ii) enforce the Terms, including investigation of potential violations hereof, (iii) detect, prevent, or otherwise address fraud, security or technical issues, (iv) respond to user support requests, or (v) protect the rights, property or safety of Twitter, its users and the public.
>
> Tip: Twitter does not disclose personally identifying information to third parties except in accordance with our Privacy Policy.
>
> […]

The Twitter Rules deal with two broad sets of issues: content boundaries, and abuse and spam.

With respect to issues related to violence against women, the following provisions of the Twitter rules are relevant:

> Content Boundaries and Use of Twitter
>
> In order to provide the Twitter service and the ability to communicate and stay connected with others, there are some limitations on the type of content that can be published with Twitter. These limitations comply with legal requirements and make Twitter a better experience for all. We may need to change these rules from time to time and reserve the right to do so. Please check back here to see the latest.

- Impersonation: You may not impersonate others through the Twitter service in a manner that does or is intended to mislead, confuse, or deceive others.

- Trademark: We reserve the right to reclaim usernames on behalf of businesses or individuals that hold legal claim or trademark on those usernames. Accounts using business names and/or logos to mislead others may be permanently suspended.

- Private information: You may not publish or post other people's private and confidential information, such as credit card numbers, street address or Social Security/National Identity numbers, without their express authorization and permission.

- Violence and Threats: You may not publish or post direct, specific threats of violence against others.

- Copyright: We will respond to clear and complete notices of alleged copyright infringement. Our copyright procedures are set forth in the Terms of Service.

- Unlawful Use: You may not use our service for any unlawful purposes or in furtherance of illegal activities. International users agree to comply with all local laws regarding online conduct and acceptable content.

- Misuse of Twitter Badges: You may not use badges, such as but not limited to the Promoted or Verified Twitter badge, unless provided by Twitter. Accounts using these badges as part of profile photos, header photos, background images, or in a way that falsely implies affiliation with Twitter may be suspended.

Abuse and Spam

Twitter strives to protect its users from abuse and spam. User abuse and technical abuse are not tolerated on Twitter.com, and may result in permanent suspension. Any accounts engaging in the activities specified below may be subject to permanent suspension.

- Serial Accounts: You may not create multiple accounts for disruptive or abusive purposes, or with overlapping use cases. Mass account creation may result in suspension of all related accounts. Please note that any violation of the Twitter Rules is cause for permanent suspension of all accounts.

- Targeted Abuse: You may not engage in targeted abuse or harassment. Some of the factors that we take into account when determining what conduct is considered to be targeted abuse or harassment are:

    - if you are sending messages to a user from multiple accounts;

    - if the sole purpose of your account is to send abusive messages to others;

    - if the reported behavior is one-sided or includes threats

[…]

Your account may be suspended for Terms of Service violations if any of the above is true.

[…]

Accounts engaging in any of these behaviors may be investigated for abuse. Accounts under investigation may be removed from Search for quality. Twitter reserves the right to immediately terminate your account without further notice in the event that, in its judgment, you violate these Rules or the Terms of Service.

We may revise these Rules from time to time; the most current version will always be at twitter.com/rules.

Important to note is that there is no specific reference to violence against women, speech which may amount to hate speech, or use of the platform to harass, abuse or intimidate on the basis of gender. In the absence of specific language to that effect, such behaviour is most likely covered by either Violence and Threats, or Targeted Abuse. Both headings in the Twitter Rules are hyperlinked to a page called Abusive Behavior Policy, which is reproduced here in full:

Abusive behavior policy

If you need to report abusive behavior to Twitter, please file a report here.

If you believe you may be in danger, please contact your local law enforcement authority in addition to reporting the content to Twitter so that the situation can be addressed both online and offline.

User disputes and false statements

Twitter provides a global communication platform which encompasses a variety of users with different voices, ideas and perspectives. As a policy, we do not mediate content or intervene in disputes between users.

Threats and abuse

Users may not make direct, specific threats of violence against others; targeted abuse or harassment is also a violation of the Twitter Rules and Terms of Service.

> For frequently asked questions about reporting abusive behavior on Twitter, click here. To learn more about what you can do when you encounter abusive behavior on Twitter and other websites, click here.
>
> Offensive content
>
> Users are allowed to post content, including potentially inflammatory content, provided they do not violate the Twitter Rules and Terms of Service. Twitter does not screen content and does not remove potentially offensive content unless such content is in violation of the Twitter Rules and Terms of Service.
>
> If you believe the content or behavior you are reporting is prohibited in your local jurisdiction, please contact your local authorities so they can accurately assess the content or behavior for possible violations of local law. If Twitter is contacted directly by law enforcement, we can work with them and provide assistance for their investigation as well as guidance around possible options. You can point local law enforcement to our Law Enforcement Guidelines.

Twitter's approach to abusive behaviour is clearly a hands-off one. They refer users to the Rules and Terms of Service and encourage users to contact local law enforcement with any serious complaints. However, they do provide for a reporting process, which is documented in the following section.

3.2. What is the process for reporting violations?

Twitter adopts a self-reporting model, and notes in the Twitter Rules that it does not actively monitor and will not censor content, except in prescribed situations. The different means of reporting a variety of violations on Twitter are all described on a page entitled How to report violations.

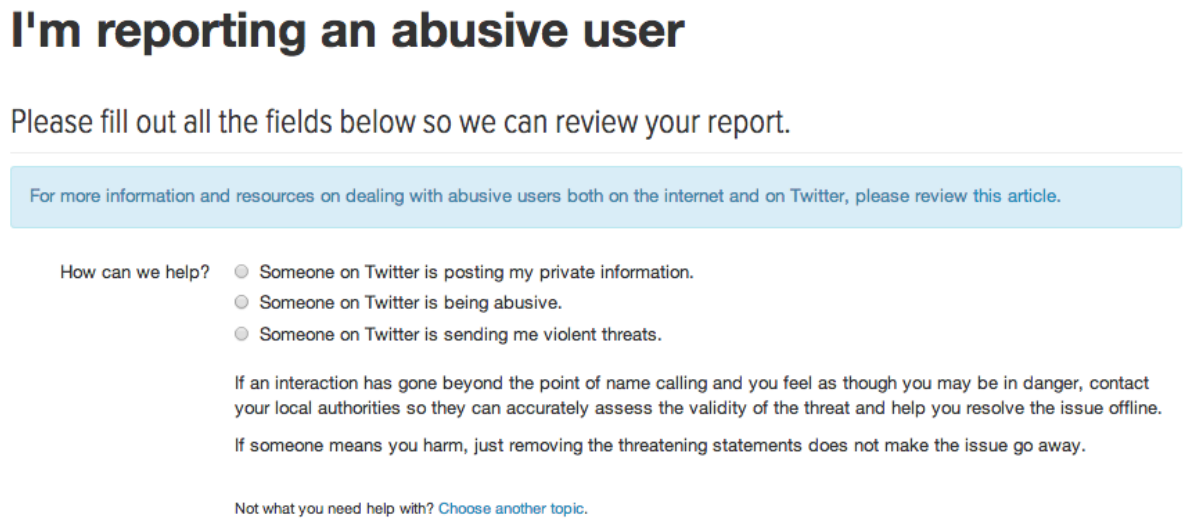
Whereas Twitter provides sufficient information to users about how to make a report, there is almost no information available as to how reports are handled, what time period applies to their resolution, who is responsible for moderating complaints and what specific standards or tests will be applied, or how users will be notified of the outcome of complaints. The only information we were able to find in this regard was an interview that Del Harvey, the senior director of trust and safety, gave to The New Yorker magazine. Harvey noted, inter alia:

- Twitter examines the clarity of the threat; if a user makes a specifically violent threat, their tweet may be removed, as may the user.
- Accounts that exist only to promote hate will be removed.
- However, if two accounts levy abuse at each other equally, it will be considered bad behaviour more than abuse.

- There are two teams of a few dozen people, based in Dublin and San Francisco, monitoring abuse and examining whether a given exchange is a two-way dialogue, whether the tweet could be construed as some kind of in-joke, whether the account is engaging only in this abusive behaviour, and whether the victim has blocked the abuser. (The team also looks at how frequently the abuser has been blocked by other users.)
- Some of the thresholds/tests that Twitter might apply in assessing reports include: how many people have blocked this user, how many have reported the user, is it content that is starting a dialogue? What was the intent of the content?[11]

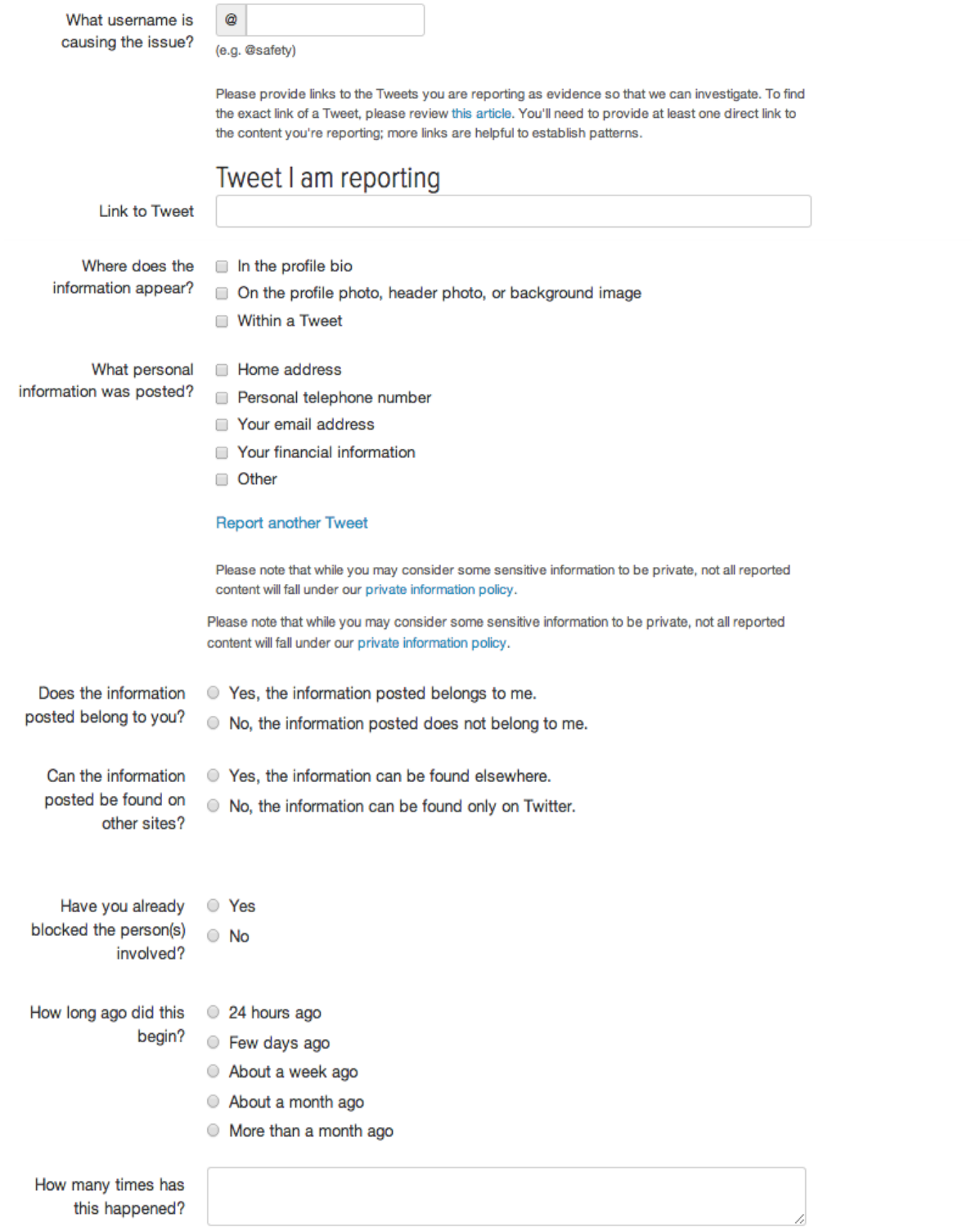
Reporting violations via the online form

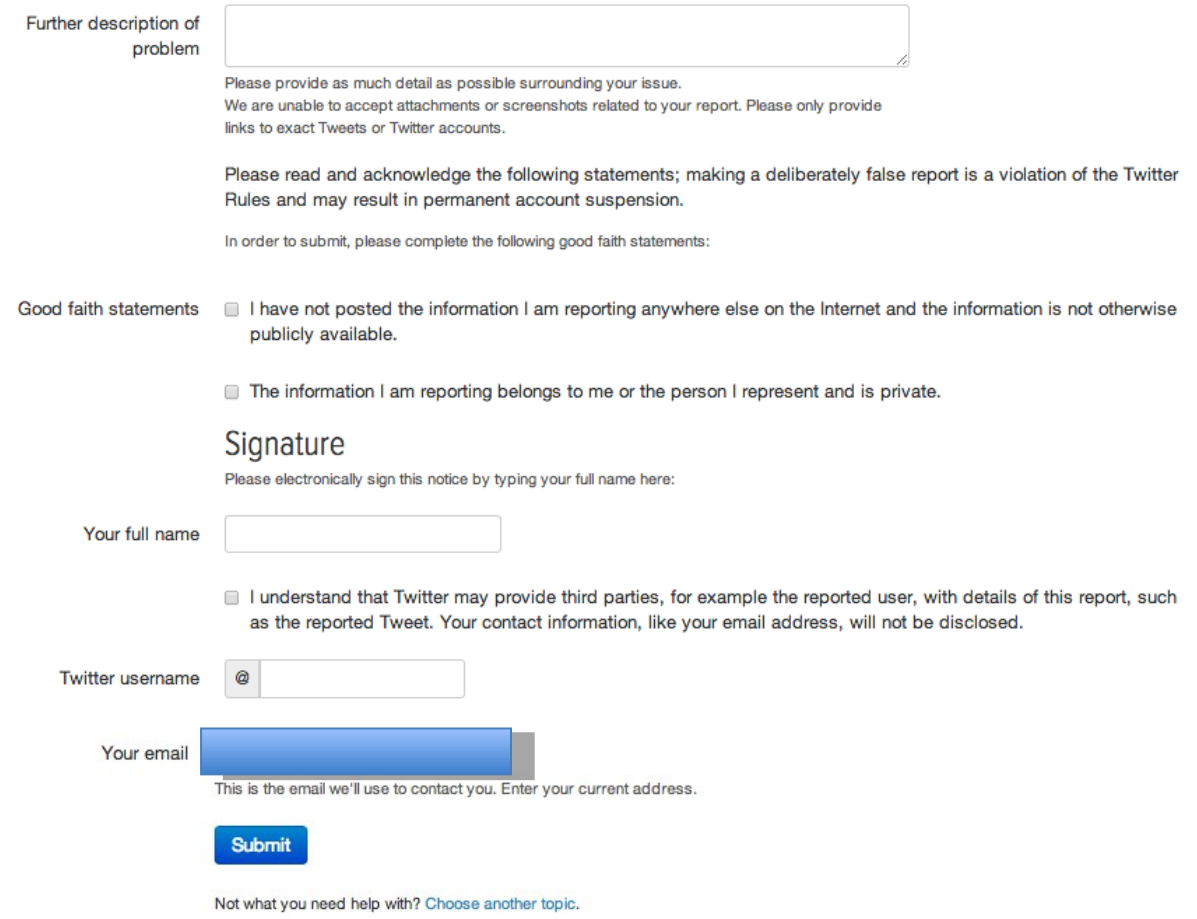
With respect to behaviour that would fall within the confines of violence against women, users can reach the reporting mechanism via a number of paths in the Twitter Help Center, including the How to report violations page or the Abusive Behaviour Policy page, which is linked directly to the Twitter Rules. Clicking through from either page takes a user to a form entitled I'm reporting an abusive user. There, users have three options to choose from:



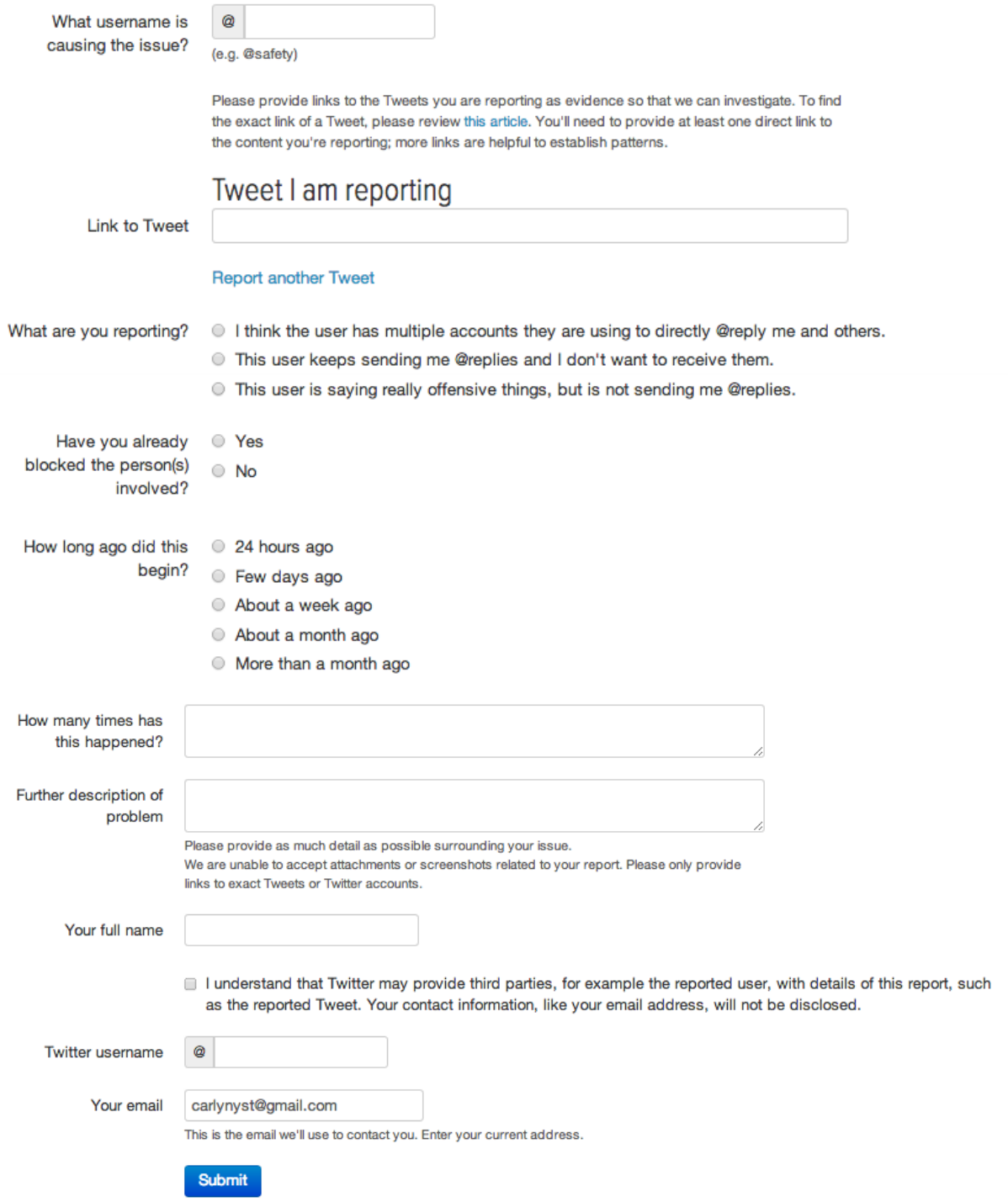
[11] Greenhouse, E. (2013, August 1). Op. cit.

When users select "**Someone on Twitter is posting my private information**" they are prompted to answer the following additional questions

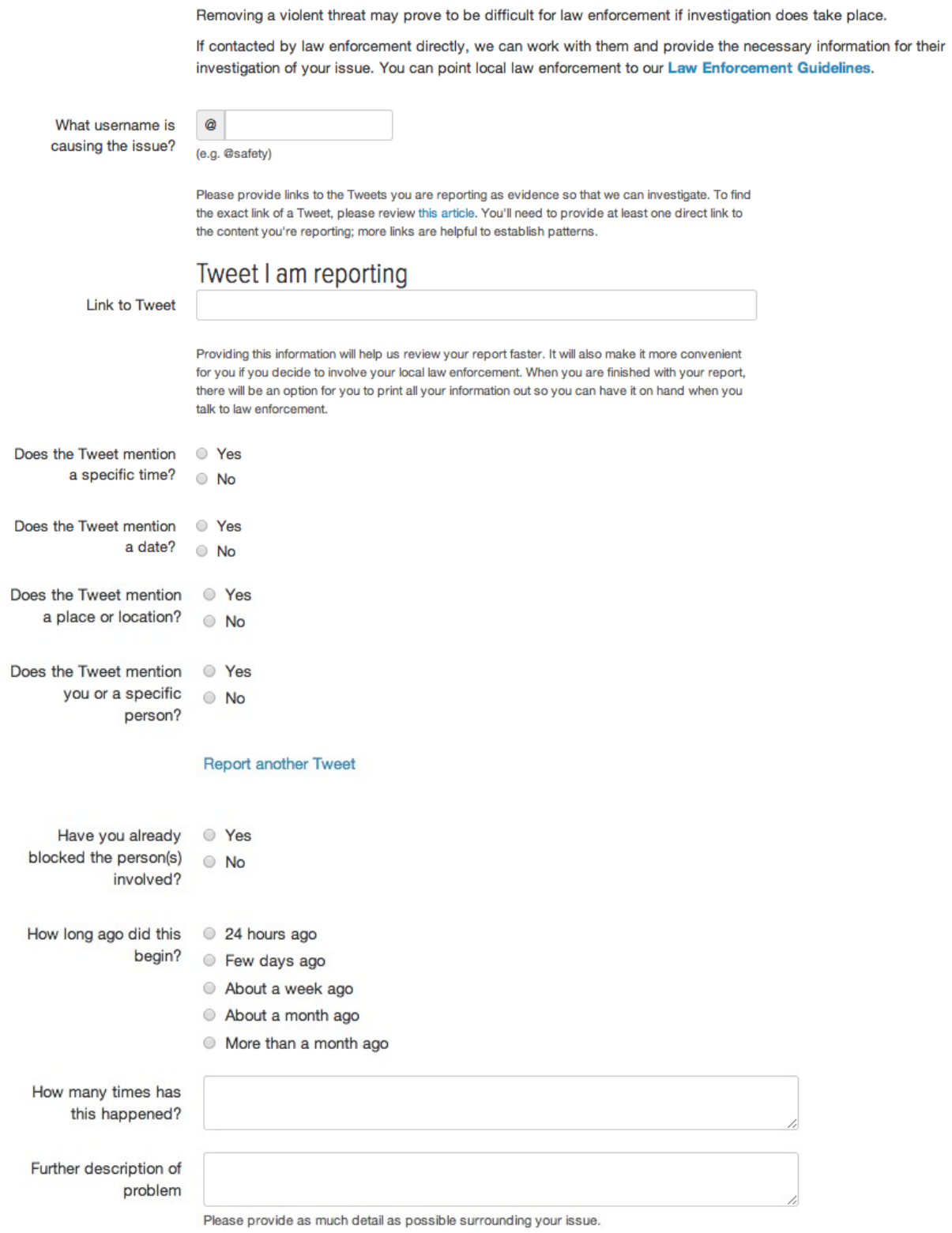


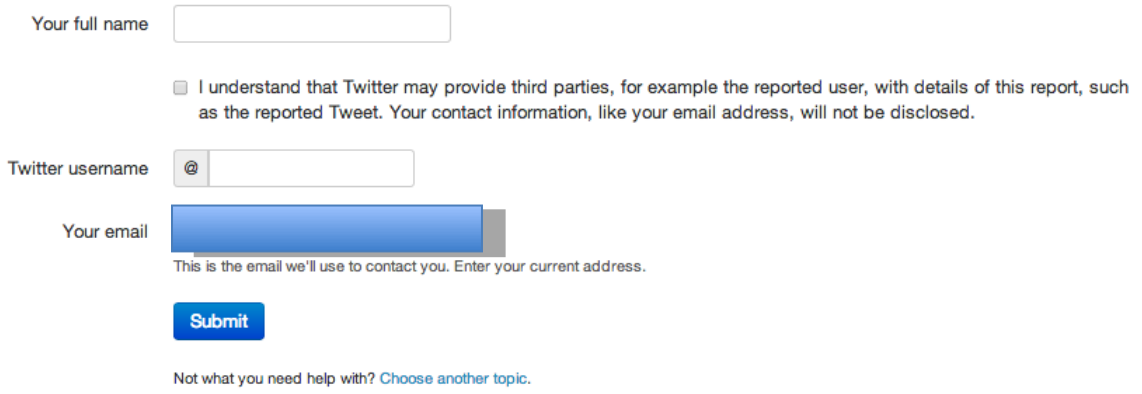

Further description of problem
Please provide as much detail as possible surrounding your issue.
We are unable to accept attachments or screenshots related to your report. Please only provide links to exact Tweets or Twitter accounts.
Please read and acknowledge the following statements; making a deliberately false report is a violation of the Twitter Rules and may result in permanent account suspension.
In order to submit, please complete the following good faith statements:
Good faith statements
I have not posted the information I am reporting anywhere else on the Internet and the information is not otherwise publicly available.
The information I am reporting belongs to me or the person I represent and is private.
Signature
Please electronically sign this notice by typing your full name here:
Your full name
I understand that Twitter may provide third parties, for example the reported user, with details of this report, such as the reported Tweet. Your contact information, like your email address, will not be disclosed.
Twitter username
@
Your email
This is the email we'll use to contact you. Enter your current address.
Submit
Not what you need help with? Choose another topic.

When users select "**Someone on Twitter is being abusive**" they are prompted to answer the following additional questions:



When users select "**Someone on Twitter is sending me violent threats**" they are prompted to answer the following additional questions:





The frequently asked questions page on Reporting abusive behaviour provides the following

> Who can report abusive behavior on Twitter?
>
> In order to investigate reports of abusive behaviors, violent threats or a breach of privacy, we need to be in contact with the actual person affected or their authorized representative. We are unable to respond to requests from uninvolved parties regarding those issues to mitigate the likelihood of false or unauthorized reports. If you are not an authorized representative but you are in contact with the individual, encourage the individual to file a report through our forms.
>
> Why can't Twitter block a user from making new accounts?
>
> IP blocking is generally ineffective at stopping unwanted behavior, and may falsely prevent legitimate users from accessing our service.
>
> IP addresses are commonly shared by numerous different users in a variety of locations, meaning that blocking a single IP may prevent a large number of unconnected users from logging into Twitter. In addition, IP addresses are easy to change and blocks can be easily circumvented by logging in from a different location, a third-party service, or one of many free websites or applications.
>
> Can Twitter give me another user's information?
>
> Per our Privacy Policy, Twitter does not release user information except as required by valid legal process. If you are working with the police or your lawyer, they will be able to help you with the appropriate correct legal process for obtaining such information. If Twitter is contacted directly by law enforcement, we can work with them and provide assistance for their investigation. You can point local law enforcement to our Guidelines for Law Enforcement.

What should I do if I receive a violent threat?

If someone has Tweeted a violent threat that you feel is credible, contact law enforcement so they can accurately assess the validity of the threat. Websites do not have the ability to investigate and assess a threat, bring charges or prosecute individuals. If contacted by law enforcement directly, we can work with them and provide the necessary information for their investigation of your issue. Click here to report a violent threat.

How do I file a report that someone is tweeting abusive messages?

You can file a report that someone is posting abusive messages by going to this page.

What happens when Twitter receives a valid report?

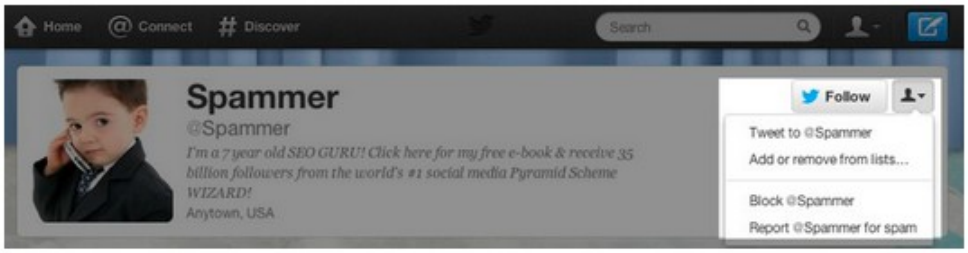
Once you have submitted your report, we will review the reported account, including the links to Tweets you'd like us to investigate. If the account is in violation of our policies, we will take action, ranging from warning the user up to permanently suspending the account.

Other ways of addressing violations

Twitter provides three other means that victims of violence against women could use to address violations on Twitter. The first is to block the user in question. This prevents the reportee from adding an individual's Twitter account to their lists, having their @replies or mentions show in the individual's mentions tab, following an individual or seeing their profile picture on the reportee's timeline. No notification is sent to the reportee that an individual has blocked them.

To block a Twitter user:

1. **Log in** to your Twitter account.
2. **Go to the profile page** of the person you wish to block.
3. **Click the person icon** on their profile page. This brings up a drop-down actions menu.
4. Select **Block** from the options listed.



The other mechanisms allow an individual to flag either media or individual tweets or direct messages for violations. Both processes are initiated at the origin of the tweet, rather than by going through the Twitter Help Center. However, they are not available on any of the Twitter clients or applications, only on the web-based version of the platform. Flagging individual tweets was initially only available for iOS (Apple-based systems) and mobile Twitter for smartphone browsers. However, reports from 2013 establish that Twitter is rolling out the feature for Android and web. This may restrict the ability of some users to make use of this function.

Twitter provides the following guidance for Reporting a tweet or direct message for violations:

> Reporting a Tweet or direct message for violations
>
> You can report Tweets or direct messages that are in violation of the Twitter Rules or our Terms of Service. This includes spam, harassment, impersonation, copyright, or trademark violations. You can report any Tweet on Twitter, including Tweets in your home timeline, the Connect or Discover tabs, or in Twitter Search.
>
> To report a Tweet:
>
> 1. Navigate to the Tweet you'd like to report.
> 2. Tap the ••• icon (or the ⋮ icon on Android) to bring up the off-screen menu.
> 3. Select **Report Tweet** and then one of the options below (Spam, Compromised, Abusive, or Block account).
> 4. Select **Submit** (or Next if reporting abuse; see below for details) or **Cancel** to complete the report or block the user.
>
> To report a direct message:
>
> 1. Select the direct message you'd like to report.
> 2. Tap and hold the message and select **Report** from the options that appear.
> 3. You can choose the option to **Report spam** or **Mark as abusive**.
>
> Reporting options:
>
> - **Spam:** this is the best option for reporting users who are using spam tactics. Please reference the Twitter Rules for information about some common spam techniques,

which include mass creation of accounts for abusive purposes, following a large number of users in a short time, and sending large numbers of unsolicited @replies.

- **Compromised:** if you think the user's account has been compromised, and they are no longer in control of their account, select this option, and we will follow up with them to reset their password and/or take other appropriate actions.

- **Abusive:** for other types of violations, including harassment, copyright or trademark violations, and impersonation, select this option.

- **Block account:** instead of reporting a user, you can select this option to block the user. Learn more about what blocking means.

What happens when I report a Tweet or direct message?

- The tweet or message will disappear from your timeline.

- Reporting a Tweet does not automatically result in the user being suspended.

What's the difference between reporting a user and reporting a Tweet or direct message?

Reporting a Tweet or direct message allows you to indicate what specific Tweet or direct message you think is in violation of the Twitter Rules or policies. If the user is violating Twitter policies without posting a Tweet or direct message (for example, mass following large numbers of users), then you should report the user as spam.

3.3. What are the support mechanisms in place for victims/survivors?

The Twitter Help Center provides the following advice for victims of "online abuse":

> Online abuse
>
> Being the target of online abuse is not easy to deal with. Knowing the appropriate steps to take to address your situation can help you through the process.
>
> When to report it?
>
> We've all seen something on the Internet we disagree with or have received unwanted communication. Such behavior does not necessarily constitute online abuse. If you see or receive an @reply you don't like, unfollow and end any communication with that user.
>
> If the behavior continues, it is recommend that you block the user. Blocking will prevent that person from following you or seeing your profile picture on their profile page or in their timeline; additionally, their @replies or mentions will not show in your mentions tab (although these Tweets may still appear in search).
>
> Abusive users often lose interest once they realize that you will not respond. If the user in question is a friend, try addressing the issue offline. If you have had a misunderstanding, it may be possible to clear the matter up face to face or with the help of a trusted individual.
>
> If you continue receiving unwanted, targeted and continuous @replies on Twitter, and feel it constitutes online abuse, consider reporting the behavior to Twitter here.
>
> Take threats seriously
>
> If you believe you are in physical danger, contact the local law enforcement authorities who have the tools to address the issue.
>
> If you decide to work with law enforcement, make sure to do the following:
>
> - document the violent or abusive messages with print-outs or screenshots
> - be as specific as possible about why you are concerned
> - provide any context you have around who you believe might be involved, such as evidence of abusive behavior found on other websites

- provide any information regarding previous threats you may have received

You can report the content to Twitter here.

Reach out to the people you trust

When dealing with negative or hurtful interactions, it can help to turn to family and friends for support and advice. Oftentimes, talking with your relatives or a close friend may help you figure out how you want to handle the situation or let you express your feelings so you can move on. There are many online resources that can help, too:

- Stop Bullying | @stopbullyinggov
- National Crime Prevention Center on Cyberbullying
- Cyberbullying Research Center
- Connect Safely | @connectsafely
- UK's Safer Internet Centre | @UK_SIC
- Anti-Bullying Pro | @antibullyingpro
- National Society for the Prevention of Cruelty to Children | @NSPCC
- The Cybersmile Foundation | @CybersmileHQ
- Pantallas Amigas | @PantallasAmigas

Help others

Trying to figure out how to help someone in such a situation can be daunting. This Twitter Support article offers some suggestions.

If you see a violent or abusive message directed at someone else, communicate your concern to the recipient and encourage them to contact Twitter and their local authorities.

The Twitter Support article referred to above reads as follows:

> Helping a friend or family member with online abuse
>
> Being the target of online abuse is not easy. If someone you know if being affected by online abuse, here are some ways that might help, or make the situation easier for them.
>
> Try to understand
>
> Just because the abuse is happening online, doesn't make it any less real. If a friend or loved one seeks your help with an abusive online situation, listen to what they have to say and take their situation seriously.
>
> Encourage them to get help
>
> While it's important to be there for them, encourage them to get professional help, whether it be a counselor, therapist, lawyer, law enforcement, or other trusted individual. If you know the individual online only, suggest they seek people offline they can talk with.
>
> Don't be a bystander
>
> If you see someone being abused online, don't look the other way. While it can be tempting to retaliate against the abuser with hurtful words, this is usually what they want you to do. Instead, reach out to your friend or family member with positive words to remove the attention from the abuser.
>
> Encourage them to report
>
> Twitter can only accept reports from the individual directly involved in the abusive situation, or their legal representation. Direct your friend or family member to file a report. We encourage people to file reports of abuse so that we can investigate the situation and take action if necessary.

3.4. At what point do intermediaries collaborate with others to facilitate access to justice?

Twitter states repeatedly throughout the content posted in the Twitter Help Center that it will assist law enforcement when contacted by them, and repeatedly encourages users to take complaints to law enforcement rather than relying on Twitter to conduct investigations.

Twitter provides Guidelines for Law Enforcement. Because of the length of that document we've omitted to include it in full, but some relevant provisions are as follows:

> Guidelines for Law Enforcement
>
> These guidelines are intended for law enforcement personnel seeking to request information about Twitter users. Information regarding requests to withhold content is available on our Country Withheld Content article; requests can be filed through our web form. More general information on Twitter's Rules can be found here.
>
> […]
>
> What User Information Does Twitter Have?
>
> User information is held by Twitter, Inc. in accordance with our Privacy Policy and Terms of Service. We require a subpoena, court order, or other valid legal process to disclose information about our users.
>
> Most Twitter profile information is public, so anyone can see it. A Twitter profile contains a profile photo, header photo, background image, and status updates, called Tweets. In addition, the user has the option to fill out location, a URL, and a short "bio" section about themselves for display on their public profile. Please see our Privacy Policyfor more information on the data we collect from users.
>
> […]
>
> Requests From Non-U.S. Law Enforcement
>
> U.S. law authorizes Twitter to respond to requests for user information from foreign law enforcement agencies that are issued via U.S. court either by way of a mutual legal assistance treaty ("MLAT") or a letter rogatory. It is our policy to respond to such U.S. court ordered requests when properly served.
>
> Non-U.S. law enforcement authorities may also submit requests for emergency disclosure under exigent circumstances, as outlined in the section titled "How to Make an Emergency Disclosure Request," above.
>
> Assisting a Twitter User
>
> If you are assisting a Twitter user with an investigation and want to obtain a copy of the Twitter user's non-public account information, please instruct the user to contact us directly (see below) to request his or her own information.

General Inquiries

Other general inquiries from law enforcement / government officials can be submitted through our web form.

Contact Information

You may fax Twitter, Inc., c/o Trust & Safety – Legal Policy, at: 1-415-222-9958.

Our mailing address is:

Twitter, Inc.
c/o Trust & Safety - Legal Policy
1355 Market Street Suite 900
San Francisco, CA 94103

Receipt of correspondence by any of these means is for convenience only and does not waive any objections, including the lack of jurisdiction or proper service.

Non-law enforcement requests should be sent through our regular support methods (https://support.twitter.com).

3.5. Evolution of Twitter's policies related to technology-related VAW, 2009 to 2014

Twitter is a relatively young platform, having been established in 2006, but in its short life it has encountered numerous difficulties with respect to its policies and procedures related to abusive and violent speech online.

In June 2013, Twitter lost a case[12] in the French courts in which it had been resisting demands to hand over user data on individuals accused of violating French law by espousing anti-Semitic hate speech via the platform. The speech had been propagated with the use of the hashtag #unbonjuif (#agoodjew) and was alleged to violate French law. Twitter had refused to hand over data on the identity of the individuals in question, in keeping with its general commitment to erring on the side of free and unrestrained expression.[13] The decision came only months after the Simon Wiesenthal Center published a report noting that there were more than 20,000 hate speech-related hashtags circulating in the Twitter-verse.[14]

[12]Keating, J. (2013, July 12). Twitter loses French hate speech case. *Foreign Policy*. ideas.foreignpolicy.com/posts/2013/07/12/twitter_loses_french_hate_speech_case

[13]Roberts, J. J. (2012, September 2). Twitter is a speech-loving tech company: the @Amac interview. *Gigaom*. gigaom.com/2012/09/02/twitter-is-a-speech-loving-tech-company-the-amac-interview/

[14]Dewey, C. (2013, March 14). New report alleges rampant hate speech on Twitter. *Washington Post*. www.washingtonpost.com/business/technology/new-report-alleges-rampant-hate-speech-on-twitter/2013/03/14/9631cda2-8bfd-11e2-9838-d62f083ba93f_story.html

The most serious and important changes to Twitter's approach to issues of violence against women came about in June and July 2013, when there were a number of high-profile incidents in which prominent women were subjected to a deluge of death, violence and rape threats on Twitter.[15] These women included including British feminist and journalist Caroline Criado-Perez, who was threatened with rape; Labour Party politician Stella Creasy, who after writing an op-ed in support of Criado-Perez received a tweet stating "You better watch your back… I'm gonna rape your ass at 8pm and put the video all over the internet"; and Guardian columnist Hadley Freeman, Independent columnist Grace Dent and Europe editor of Time magazine Catherine Mayer, who were subjected to bomb threat tweets.

Around the same time, feminist media critic Anita Sarkeesian tweeted, "I've reported numerous rape threats to @Twitter. This is how they respond: 'The account is currently not in violation of the Twitter Rules.'" She also posted a screen grab of one such tweet, showing the account @CoolDehLa tweeting, on December 26, 2012, "@femfreq I will rape you when i get the chance." Sarkeesian wrote, "Twitter says 'We have found the reported account is currently not in violation of the Twitter Rules at this time.'" Sarkeesian's two tweets were retweeted more than 7,000 times.[16]

Criticism of Twitter's inability to respond effectively to violence against women began to build. The incidents prompted the establishment of a Change.org petition for the introduction by Twitter of a report button to enable direct reporting of violations at the source. The petition gained more than 140,000 signatures.[17]

On 20 July 2013, Del Harvey, Twitter's head of safety, responded by publishing a blog post on the platform's site entitled "We Hear You".[18] She announced the introduction of a report button, and committed Twitter to improving the effectiveness and efficiency of its services. Around the same time, Tony Wang, general manager of Twitter UK, posted a series of tweets saying abuse was "simply not acceptable", and offering his personal apologies.[19]

The report button has been introduced, although it is not available on all platforms and services. It is unclear what further measures Twitter has taken to improve the effectiveness of its reporting mechanisms.

On 13 December 2013, Twitter reversed changes to the "blocking" mechanism[20] that it had announced only the previous day. The changes had allowed blocked users to continue viewing the tweets of and interact with accounts that had blocked them, while remaining invisible so that the reporting user could not see that the blocked user was following them. This was a serious

[15]Ensor, J. (2013, August 3). Twitter updates rules in clampdown on abuse. *The Telegraph*. www.telegraph.co.uk/technology/twitter/10220382/Twitter-updates-rules-in-clampdown-on-abuse.html

[16]Greenhouse, E. (2013, August 1). Op. cit.

[17] www.change.org/en-GB/petitions/twitter-add-a-report-abuse-button-to-tweets

[18] The blog post is no longer hosted on the Twitter site, but can be accessed at: web.archive.org/web/20130806015808/http://blog.uk.twitter.com/2013/07/we-hear-you.html

[19] Ensor, J. (2013, August 3). Twitter boss personally apologises to female victims of abuse. *The Telegraph*. www.telegraph.co.uk/technology/twitter/10220410/Twitter-boss-personally-apologises-to-female-victims-of-abuse.html

[20]BBC. (2013, December 13). Op. cit.

departure from the previous situation, in which users could prevent a blocked user from following them or viewing their tweets.

Twitter said that the changes were made so that the blocked user would not know they had been blocked, a development that often leads blocked users to retaliate against the reporting user. Twitter Chief Executive Dick Costolo said that the changes were widely requested by victims of abuse.

However, within hours of announcing the changes, Twitter was flooded with the complaints of angry users, and confronted with an online petition to reverse the change.[21] Twitter reversed its changes almost immediately, reverting to the previous policy.

[21] Shih, J. (2013, December 13). Op. cit.

ISBN: 978-92-95102-19-4
APC-201407-WRP-R-EN-DIGITAL-210