29 Nov 2019

Mahima Kaul
Director, Public Policy
Twitter India

mkaul@twitter.com
@misskaul

Dear Mr Patel,

Thank you for the email and the letter you have sent to us. Thank you for the email and the letter you have sent for us. We take safety very seriously at Twitter and would be happy to share our progress and perspective with you.

Last year, we shared that building a Twitter free of abuse, spam and other behaviours that distract from the public conversation is one of our top priorities. Since then, we've made strides in creating a healthier service and we've continued to further invest in proactive technology to positively and directly impact people's experience on the service.

Twitter is "What's Happening" across the globe — and what we've seen happening is powerful voices and movements come together to speak up for women's rights. Women and allies around the world are joining together on Twitter to share experiences, challenges and successes. They are sharing what they want to see change, fostering dialogue and debate, and amplifying their voices to new audiences. It is to protect their voices -- and the voices of all those who use our service -- that we continue to work on the safety of the service.

Product features
The power of Twitter lies in the fact that we are an open, public and real time service. Our service is reflective of real conversations happening in the world and that sometimes includes perspectives that may be offensive, controversial, and/or bigoted to others.

We have a series of tools, built into our product, to help keep people safe and give them control over what they see and who they interact with. These tools include:

- **Unfollow:** If someone wants to stop seeing a particular account's Tweets in their home timeline, they can unfollow the account. They can still view the Tweets on an as-needed basis by visiting the profile, unless the Tweets on the profile are protected.
- **Block:** People can restrict specific accounts from contacting them, seeing their Tweets, and following them by blocking the account.
- **Advanced Block:** People can export their list of blocked accounts to share with another person and import someone else's list of blocked accounts using the Advanced Block feature.
- **Mute:** People can remove an account's Tweets from their timeline without unfollowing or blocking it. They can also use Advanced Mute for particular words, conversations, phrases, usernames, emojis, or hashtags.
- **Disable Receive Direct Message setting:** People can prevent accounts that they do not follow from DMing them by disabling the Receive Direct Message setting.

- **Filtered notifications:** People can apply different filters on the types of notifications they receive. Mute Notifications allows people to mute phrases and words they'd like to avoid seeing in their notifications. Advanced Filters allows them to disable notifications from certain types of accounts or at certain time periods - for example if their account is receiving a lot of sudden attention.
- **Protected Tweets:** When a person signs up for Twitter, their Tweets are public by default which means that anyone can view and interact with them. If a person protects their Tweets, this will make their account private and other Twitter users will have to send a request if they want to follow the account.
- **Safe search:** The Safe Search function removes potentially sensitive content by default, as well as accounts people have blocked and muted from search pages.
- **Sensitive media:** People can opt out of seeing certain imagery that may be sensitive. Twitter's default setting is to place potential sensitive material behind a warning. This can be adjusted in settings.

Updates to the Twitter Rules

The Twitter Rules are a living document and we are continually working to update, refine, and improve both our enforcement and our policies. This work is informed by in-depth research around trends in online behavior both on and off Twitter, feedback from the people who use Twitter, and input from a number of external entities.

Our rules are in place to ensure all people can participate in the public conversation freely and safely. Violence, harassment and other similar types of behavior discourage people from expressing themselves, and ultimately diminish the value of global public conversation.

In June this year we undertook a major refresh of the Twitter Rules to make them simpler and easier to understand. We've gone from about 2,500 words to under 600. Each Rule is now 280 characters or less (the length of a Tweet) and describes exactly what is not allowed on Twitter.

We organised our rules around three categories — Safety, Privacy, and Authenticity — which makes it easier for people to find the information they're looking for more quickly.

Although we have simplified the language of our rules considerably, where possible, we've updated our rules pages to include more detail such as examples, step-by-step instructions about how to report, and details on what happens when we take action.

Relevant to your enquiry, in the area of Safety, our rules are as follows:

- **Violence:** You may not threaten violence against an individual or a group of people. We also prohibit the glorification of violence. Learn more about our violent threat and glorification of violence policies.
- **Child sexual exploitation:** We have zero tolerance for child sexual exploitation on Twitter. More information can be found here.
- **Abuse/harassment:** You may not engage in the targeted harassment of someone, or incite other people to do so. This includes wishing or hoping that someone experiences physical harm. More information can be found here.

- **Hateful conduct:** You may not promote violence against, threaten, or harass other people on the basis of race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease. More information can be found here.
- **Suicide or self-harm:** You may not promote or encourage suicide or self-harm. More information can be found here.
- **Sensitive media, including graphic violence and adult content:** You may not post media that is excessively gory or share violent or adult content within live video or in profile or header images. Media depicting sexual violence and/or assault is also not permitted. More information can be found here.

In the area of Privacy, our rules are as follows:

- **Private information:** You may not publish or post other people's private information (such as home phone number and address) without their express authorization and permission. We also prohibit threatening to expose private information or incentivizing others to do so. More information can be found here.
- **Non-consensual nudity**: You may not post or share intimate photos or videos of someone that were produced or distributed without their consent. More information can be found here.

As mentioned above, we are constantly reviewing and updating our rules to ensure they keep pace with the ways in which people use our service. Some of the changes we made in recent years in the area of online safety include (but are not limited to) updating the list of abusive behaviors we prohibit to include unwanted sexual advances, posting or sharing intimate photos or videos of someone that were produced or distributed without their consent, wishes or hopes of harm, and threats to expose or hack someone.

Last year we expanded hateful conduct and media policies to include abusive usernames and hateful imagery. We also updated rules around violence and physical harm to include the glorification of violence and violent extremist groups. Most recently, we updated our Hateful Conduct policy to prohibit dehumanising language on the basis of religion.

Our policies and enforcement options evolve continuously to address emerging behaviors online

Reporting and enforcement
We use a combination of human review and technology to help us enforce our rules. Our team reviews and responds to reports, 24/7; and they have the capacity to review and respond to reports in multiple languages.

Our team undergoes in-depth training on our policies, ensuring we're considering social and political nuances, and taking local context and cultures into account.

We accept reports of violations from anyone - in fact, we have a bystander policy that enables anyone who witnesses abuse and harm on our service to report these. Sometimes we also need to hear directly from the target to ensure that we have proper context.

The number of reports we receive does not impact whether or not something will be removed. However, it may help us prioritize the order in which it gets reviewed.

Reports therefore, are looked at on a case-by-case basis. Unless a violation is so egregious that we must immediately suspend an account, we first try to educate people about our Rules and give them a chance to correct their behavior. We show the violator the offending Tweet(s), explain which Rule was broken, and require them to remove the content before they can Tweet again. If someone repeatedly violates our Rules then our enforcement actions become stronger. This includes requiring violators to remove the Tweet(s) and taking additional actions like verifying account ownership and/or temporarily limiting their ability to Tweet for a set period of time. If someone continues to violate Rules beyond that point then their account may be permanently suspended.

If we identify an account or Tweet that violates the Twitter Rules, there are a range of enforcement options we may pursue. These include limiting Tweet visibility, requiring a person to delete a Tweet, placing accounts in read-only mode or, for more serious or repeat offences, permanently suspending an account. Certain types of behavior may pose serious safety and security risks and/or result in physical, emotional, and financial hardship for the people involved. These egregious violations of the Twitter Rules — such as posting violent threats, non-consensual intimate media, or content that sexually exploits children — result in the immediate and permanent suspension of an account.

Twitter account holders can appeal enforcement decisions either in app or visiting help.twitter.com/appeals. We have a specially trained, global team to evaluate appeals in line with any additional context provided by the account holder and against the Twitter Rules.

Twitter does not collect detailed data covering specific attributes of our account holders such as gender or caste. Indeed, Twitter allows account holders to remain pseudonymous, which we believe is an important protection to free expression, particularly in parts of the world where the repercussions for certain activities may put an individual at risk.

Our progress

The Twitter Transparency Report is a bi-annual highlights trends in legal requests, intellectual property-related requests, Twitter Rules enforcement and platform manipulation (amongst other things).

The latest Twitter Transparency Report, which covers the period January 1 to June 30, 2019 details the progress we have made.

As called for by Amnesty, the report now includes data broken down across a range of key policies detailing the number of reports we receive and the number of accounts we take action on.

Across Twitter, more than 50% of Tweets we took action on for abuse were proactively surfaced using technology, rather than relying on reports from people who use Twitter. **This is important progress because it's reducing the burden on those people who may be experiencing abuse and harassment to report to us.**

Over this period we saw a 105% increase in accounts actioned by Twitter (locked or suspended for violating the Twitter Rules).

With regards to specific policies, we have also made important progress. Under our Private Information Policy we saw a 48% increase in accounts reported for potential violations of our private information policies and we suspended 119% more accounts than the previous reporting period. This increase may be attributed to the launch of improvements to our reporting flow that make it easier to report private information, as well as changes to our internal enforcement processes which permit bystanders to report potential private information violations for review.

There was a 48% increase in accounts reported for potential violations of our Hateful Conduct policies. We actioned 133% more accounts compared to the last reporting period. Similarly, we saw a 22% increase in accounts reported for potential violations of our abuse policies. We took action on 68% more accounts compared to the last reporting period.

Safety and Awareness Campaigns

Twitter runs a number of public campaigns aimed at increasing awareness on online safety and helping people who use Twitter take control of their online experience. These campaigns include Tweesurfing, #PositionOfStrength, #EduTweet and partnerships with NGOs for online campaigns on safety, and workshops to upskill non profits on how to use Twitter safety and report abuse.

#PositionOfStrength, launched in India in 2016, is aimed at women on Twitter and experience staying online. Part of the focus is to help women understand how to use Twitter's tools to curate an experience that they enjoy, and also how to report to tweets on the service. The objective is to ensure that women don't cede space online because they feel unsafe, but to work together to make the space safer for them. As part of the #PositionOfStrength movement, Twitter India and our partners have **hosted six roundtables and workshops with women leaders in Delhi, Mumbai and Bangalore**, to explore increased empowerment and safety for women, both online and in the physical world. In fact, our #हमसेहैहिम्मत event in New Delhi, when CEO Jack Dorsey visited India was to showcase the vibrancy of the Indian Twitter community as part of our #PositionOfStrength series. The speakers included representatives from the National Campaign on Dalit Human Rights, @FeminismInIndia, among others.

#EduTweet is focused at educators and teachers, to teach them how to use Twitter, media literacy, and how to stay safe online so that school teachers are equipped to answer questions their students might have on online safety. The program also teaches educators on how to leverage Twitter in the classroom, and network on issues and subjects with other teachers across the globe. **650 school principals, teachers and trustees were connected and trained through this program in Mumbai, Delhi, Bangalore and Ahmedabad.**

Tweesurfing leveraged the power of people on Twitter and influencers to talk about their journey on the service and share video clips of online safety tips. Aimed at millennials, the campaign involved offline workshops at colleges

across India, and resulted in a repository of best practises on a website and Twitter feed. Tweesurfing also involved key influencers in different fields talking about how to use Twitter's unique product features to stay safe when using the service. **Over 100 influencer interviews, 4 events, 7 TweetChats, and 8 workshops were held across the country during the campaign.**

#WebWonderWomen was a collaboration with the Ministry of Women and Child Development and not for profit partner Breakthrough, to elevate the voices of women who are highlighting important issues and making a positive impact on the platform within smaller niches. They were also trained in using Twitter better including how to stay safe. **We received over 200 applications of which 30 women achievers were awarded as part of the initiative.**

Twitter supports many campaigns and events which focus on online safety and mental health; SheThePeople's Online Safety Summit, ResponsibleNetism's National Cyber Psychology Conference, mental health campaigns with White Swan Foundation, resourcing on information on CSE content with Aarambh India, E-Raksha Online Safety Summit (NCERT) and CyberPeace Corps' Cyber Safety Summit and Cyber Kumbh.

Further, we have a number of safety partners in India who help us with feedback on proposed policies and craft partnerships to talk about online safety. These include Center for Social Research, White Swan Foundation, Breakthrough, Youth Ki Awaaz, Aarambh India, among others.

Product updates

We continue to evolve our product with the intent of improving the experience for people who use our service. In recent years, some of the changes we've made from a safety perspective include (but are not limited to):

- Updating our notification service so that people suspended for abusive behaviour will be emailed with the violating content and the rule that was broken
- Providing an option for people who report violative content to us to opt-in to have reported Tweets included in receipts Twitter sends, both in-app and through email
- Updating our reporting flow to offer more detail on what Twitter defines as a 'protected category'
- Announcing that new behaviour-based signals will be used to influence how Tweets are organised and presented in areas like Search and Conversation to reduce the visibility of lower quality and unhealthy content
- Announcing the acquisition of Smyte, experts in safety, spam, and security, to help us in our efforts to improve the health of the public conversation on Twitter.
- Strengthening our enforcement of policy around chat in live video
- Launching a filter for DMs targeting low quality messages
- Updating the product so that account holders don't see Tweets they've reported, and also providing an in-timeline notice of action taken against reported Tweets

This year, we:

- Improved the reporting flow for private information policy violations. Reporters can now add additional context before submitting.
- Informed users about new in-app appeal process which allows us to get back to people who report to us 60%faster than before.
- Changed the number of accounts users can follow per day from 1,000 to 400 to combat spam - often the underlying cause of abuse and harassment.
- Launched a Public Interest Interstitial for violative Tweets that may still be of interest/value to the public.
- Announced the global roll out of the Hide Author Moderated Replies feature which gives Twitter account holders additional control over what replies are initially visible under their Tweets.

Several of these changes address concerns raised by Amnesty International previously and we are grateful for your feedback to help us improve Twitter.

Our work to build a safer, healthier Twitter product will never be done. To that end, our focus on conversational health moving forward is in three key areas:

1. Dynamics: Making people feel safer and more comfortable talking on Twitter is part perception and part control. We're making deliberate decisions around Tweet visibility and extending that decision making power to people who start conversations. An example of this is Hide Replies (as outlined above).
2. Incentives: We want to encourage people to have healthier conversations by providing more context and more nuanced ways for people to express themselves. We are going to work on this through a series of tests and new features because we know there is no silver bullet. For example, we will be revisiting the engagement options we provide and how they work (e.g., the like, retweet, retweet with comment).
3. Comprehension: We want to make conversations on Twitter better - it should be easier to understand what's being said and who's saying it. We also want to emphasize what people say/the conversation itself, rather than how many likes it has. For example, we have been testing a new design for conversations on Twitter to help clarify these areas. We plan to roll this new design out to users in 2020.

More information on some of the product fixes and experiments we are running in this space can be found here.

Indian General Elections, 2019

Improving the collective health of the public conversation is a top priority for our company, and protecting the integrity of elections is an essential part of our mission. A summary of some of our work to protect the health of the public conversation around the 2019 Indian General Elections can be found here.

As outlined in the blog, the approach to elections at Twitter is comprehensive, cross functional and bespoke to our platform. For the 2019 Indian General Elections, we focused on seven key areas:

1. Evolving our product
2. Updating the Twitter Rules
3. Addressing manipulation

4. Scaling our internal team
5. Improving language support and cultural context
6. Working with political parties and election officials in India
7. Serving the public conversation

Using our proprietary-built internal tools, the team proactively protects trends, supports partner escalations, and identifies potential threats from malicious actors.

In the area of online safety, we introduced a partner feedback portal to a number of civil society partners to escalate issues on Twitter to us. We worked with the Election Commission of India and adopted a voluntary code of ethics along with other social media companies to highlight our commitment to serve the public conversation.

After the elections, we have launched a new campaign; #HerPoliticalJourney to celebrate the struggle, triumph and indomitable spirit of women politicians. Up to 19 women politicians participated in the campaign, which seeks to share the stories of women on Twitter. We believe that along with our work on making the service safer, conversations by women will also encourage more women to come out and use Twitter to their benefit.

Verification

We announced late 2017 and in an updated Tweet in 2018 that our public verification process is currently closed. However we do still work to verify public figures on a case by case basis. For example, working with political parties to verify candidates, elected officials, and relevant party officials around the time of elections. We verify these accounts to empower healthy conversations, and to provide confidence that these public figures are whom they claim to be.

We have one global set of Rules for the hundreds of millions of people who use Twitter and we enforce these Rules judiciously and impartially.

Engaging with Twitter users

Policy feedback

Our teams routinely meet representatives from civil society, academics, journalists, government stakeholders, influencers and other people active in the public conversation on Twitter. These meetings help us better understand the experience of people using our service.

We also host sessions with our senior executives when they visit India, to help them better understand how Twitter is used in India and to enable them to receive direct feedback from different groups, including marginalized groups, who are active on Twitter. This feedback is then funneled back into the company and influences the policy and product changes we enact.

We also have started opening some of our proposed policy changes to public comments, including our policy on dehumanized content, and our policy on synthetic and manipulated media. This feedback process is to ensure we consider global perspectives and how our policies may impact different communities and cultures.

Content moderation at Twitter

When it comes to content moderation, we use a combination of human review and technology to help us enforce our rules.

Human Review
Our team reviews and responds to reports, 24/7. Our focus is on ensuring we are covering the most widely used languages on Twitter in each market. Our global team undergoes in-depth training into our policies, and we also have an intensive focus on local language, culture, and context, ensuring we're taking social and political nuances into account. For example, we have native language speakers in major Indic languages used on Twitter.

One of the underlying features of our approach is that we take a behavior-first approach. That is to say, we look at how accounts behave before we look at the content they are posting. Context matters. When determining whether to take enforcement action, we may consider a number of factors, including (but not limited to) whether:

- the behavior is directed at an individual, group, or protected category of people;
- the report has been filed by the target of the abuse or a bystander;
- the user has a history of violating our policies;
- the severity of the violation;
- the content may be a topic of legitimate public interest

The protection of our employees - regardless of where they operate - is central to our company values. Our dedicated employee assistance programs are designed to ensure our teams feel safe, secure, and respected in their work. We have built a diverse range of on-site services, including regular on-site counseling, training, and person-to-person support, particularly for those whose work may involve reviewing sensitive content. We regularly audit our support offerings to ensure they are fit for purpose and meet our global standards.

Technology
We believe that long-term success requires moving from manual, report-based services, to automated, proactive services - whereby every process, workflow and support scenario moves through a lifecycle of manual to automation, benefiting a continuous improvement mindset. The more we can leverage these to minimise the exposure to content, the less frequently our employees and contractors will come into contact with it.

Therefore, we proactively enforce policies and use technology to halt the spread of content propagated through manipulative tactics, such as automation or attempting to deliberately game trending topics.

Our Site Integrity team is dedicated to identifying and investigating suspected platform manipulation on Twitter, including activity associated with coordinated malicious activity that we are able to reliably associate with state-affiliated actors. In partnership with teams across the company, we

employ a range of open-source and proprietary signals and tools to identify when attempted coordinated manipulation may be taking place, as well as the actors responsible for it. We also partner closely with governments, law enforcement, academics, researchers, and our peer companies to improve our understanding of the actors involved in information operations and develop a holistic strategy for addressing them.

In the first six months of 2019, we challenged more than 97 million suspected spam accounts.

As many of the actors engaged in this activity take steps to obfuscate their location, we do not believe that it is possible to produce a robust breakdown of this data on a by-country basis.

It's important to note that the way we approach content moderation at Twitter is bespoke to our platform - it works for Twitter first and foremost. As an open service with hundreds of millions of Tweets shared daily, technology is critical to our ability to respond at scale.

We empower people to understand different sides of an issue and encourage dissenting opinions and viewpoints to be discussed openly. This approach allows many forms of speech to exist on our platform and, in particular, promotes counterspeech: speech that presents facts to correct misstatements or misperceptions, points out hypocrisy or contradictions, warns of offline or online consequences, denounces hateful or dangerous speech, or helps change minds and disarm.

Requests from Law Enforcement
Twitter has dedicated contact channels for law enforcement and we respond to legal process issued in compliance with applicable law. More information can be found here.

Twitter is committed to working with governments around the world to encourage healthy behavior on the service. We are in regular contact with Indian law enforcement officials.

**

We are dedicated to making Twitter a safe place for free expression. On Twitter, everyone should feel safe expressing their unique point of view with every Tweet – and it's our job to make that happen.

While updating our products, policies, and processes is critical, we believe addressing the broader challenge of safety for women online requires collaboration between governments, civil society, and NGOs. In this regard, we would be pleased to work with Amnesty India towards a common goal of making the internet a safer space for women.

Sincerely,

Mahima Kaul
Director, Public Policy
Twitter India