By The Numbers

What Content Social Media Removes And Why

By Malena Dailey

NetChoice

Executive Summary

Social media is all about connecting with and learning from others. Whether you use Facebook to connect with family and old friends, use Twitter to keep up on current events, or use TikTok to learn about the latest trend, social media should be a positive and uplifting experience. Nobody wants their time online dominated by hateful, dangerous, or illegal posts.

Content moderation allows online platforms to remain safe and open avenues for people of all backgrounds to connect and create. Far from being something that diminishes speech, it is the very thing that makes users feel safe and confident enough to express themselves online.

6 BILLION POSTS were removed in the second half of 2020.

Social media sites are clearly working hard to remove huge swathes of harmful content online.

Total Posts Removed For Specific Community Guidelines Violations

(July-December 2020)

	Graphic Violence	Child Sexual Exploitation	Hateful Content	Abuse Or Harassment	Spam	All Community Guidelines Violations
Total Posts Removed	45,180,656	51,796,162	65,633,026	25,053,853	2,915,179,425	5,903,919,504

That means users were protected against spam, defamatory content, illegal material, bullying, and harassment 6 billion times. That requires the constant effort and vigilance of online platforms to take down harmful content before it is even seen, like scams, graphic violence, and child sexual exploitation.

While content moderation is vital to social media's continued success, we recognize that how Community Guidelines operate, what types of content is targeted for removal, and the enforcement methods used by different companies can cause confusion. This report will use data from July to December 2020 to examine and clarify the ways that social media platforms keep their users safe and keep the internet an awesome tool for creativity and free expression.

Violations of Platform Rules and Community Guidelines Globally

When creating an account on social media platforms, users are asked to accept a company's designated guidelines regarding what types of accounts and posts are allowed on their sites. By creating an account and accepting a platform's terms of service, users who do not comply face consequences imposed by the platform, the most common of which being removal. Posts and accounts which violate these are removed according to each company's enforcement procedures, discussed later in this report. In the second half of 2020, platforms took down the following number of accounts and posts for violations of their Community Guidelines.

Posts Taken Down for Community Guideline Violations
(July-December 2020)

	Twitter	Facebook	Instagram	TikTok	YouTube	Pinterest	Snap
Total Posts Taken Down	4,470,600	5,720,200,000	65,277,800	89,132,938	17,194,632	2,100,253	5,543,281

Accounts Suspended for Community Guideline Violations
(July-December 2020)

	Twitter	TikTok	YouTube	Pinterest	Snap
Accounts Suspended	1,009,083	6,144,040	3,859,685	4,269	47,558

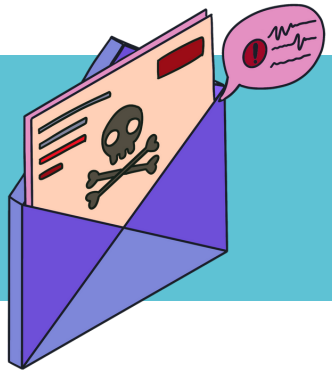
While there are differences between the policies of individual companies, objectionable and illegal material are prohibited by all and are removed when detected. Particularly egregious violations such as the spread of child sexual abuse material are both removed and reported to law enforcement and the National Center for Missing and Exploited Children (NCMEC). Some of the most common reasons for removal of a post or account include harm to the safety of minors, graphic depictions or threats of violence, and harassment.

Number of Posts Removed by Rule Violation
(July-December 2020)

	Twitter	Facebook	Instagram	TikTok	YouTube	Pinterest	Snap
Graphic Violence	59,933	34,800,000	9,700,000	267,398	13,861	1,754	337,710
Child Sexual Exploitation/ Minor Safety	9,178	17,800,000	1,809,400	32,087,857	40,383	1,794	47,550 (accounts deleted for CSAM specifically)
Hateful Content	1,628,281	49,000,000	13,100,000	1,782,658	42,013	2,487	77,587
Abuse Or Harassment	1,448,418	9,800,000	7,600,000	5,882,773	79,902	3,763	238,997

Platforms require the ability to moderate content in order to ensure that users' feeds are not flooded with this type of harmful content. By prioritizing removal of these types of posts, platforms can foster an online environment where users feel safe to browse their sites without risk of coming across graphic photos and videos without warning.

Spam



Spam is the inauthentic content or accounts, often spread in bulk through excessive posting.

Social media companies also moderate content to prevent platforms from being overrun by spam. Spam can pose risks to users if spread through the sharing of links which, when opened, can expose users to malware or online scams. A significant portion of content removed by platforms is for violations of spam policies.

On Facebook alone, nearly half of the 5.7 billion posts removed in the second half of 2020 were considered spam.

Number of Posts Removed for Violation of Spam Policies
(July-December 2020)

	Graphic Violence	Child Sexual Exploitation	Hateful Content	Abuse Or Harassment	Spam	All Community Guidelines Violations
Total Posts Removed	45,180,656	51,796,162	65,633,026	25,053,853	2,915,179,425	**5,903,919,504**

*Number of instances of reported spam on Twitter, which result in post removal, account removal, or permanent suspension depending on the severity of the violation.

Proactive Removal

Proactive removal aims to identify and take down content which violates Community Guidelines as quickly as possible in order to limit the amount of users exposed to a harmful post.

Platforms enable users to report posts and accounts, but social media companies acknowledge that taking down content is not sufficient if the content in question reaches a wide audience prior to removal. Because of this, companies report not only the number of posts removed but also the degree to which they were seen prior to removal. While platforms enable users to report posts and accounts, the aim of limiting exposure requires that a combination of artificial intelligence and human reviewers take down violating content as quickly as possible after it was posted.

This proactive approach to removal mitigates the threat of spam, violent, or otherwise offensive content being spread beyond the account responsible for the original post.

Twitter

Of the 4.5 million Tweets removed from July-December 2020, **77% were removed proactively** by Twitter, receiving less than 100 views before removal. **Only 6% of removed Tweets had over 1,000 views** when removed by the platform.

Facebook

Facebook defines their “proactive rate” as being the amount of content and accounts acted on that were found or flagged by Facebook prior to being reported by users. Across all categories of Community Standards infringement, Facebook had an **average proactive rate of 89%.** The proactive rate for severe violations such as child sexual exploitation, terrorist organizations, and violent or graphic content were all 99-100%.

Instagram

Instagram uses the “proactive rate” used by Facebook to describe the amount of content and accounts flagged by Instagram prior to being reported by users for violations of the platform's Community Standards. Of all prohibited content, Instagram acted on 86% before being reported by users.

TikTok

Of the 89.1 million videos removed from TikTok for violations of Community Guidelines, 92.4% were identified and removed by the platform before being reported by users. **83.3% were removed before receiving any views**, and 93.5% were removed within 24 hours of being posted.

YouTube

Of the 17.2 million videos removed by YouTube in the second half of 2020, approximately **74.1% were removed before receiving 10 or more views.** 40% of videos which violated YouTube’s Community Guidelines were removed from the platform before receiving any views.

Pinterest

Across all categories within Pinterest’s Community Guidelines, 91% of violating content was removed before receiving 100 views. **68% of violations were removed before receiving any views.**

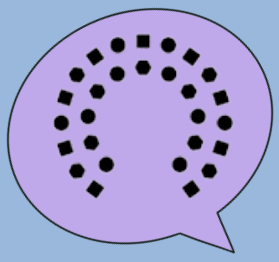
Snapchat

Snapchat’s Violative View Rate represents the percentage of views of Snapchat content which contains violations of their community guidelines. In the second half of 2020, the Violative View Rate was 0.08%, meaning **for every 10,000 views, only 8 contained content that violated Community Guidelines**.

Transparency in Moderation Decisions

Social media companies recognize the need for transparency when moderating content to ensure that users know how to safely and effectively use their platforms.

In order to approach this successfully, all the companies included here publish periodic transparency reports to provide users with data about the type of content being taken down. Platforms also include appeals processes in their sites to allow for further review of violative content, empowering users to challenge the decisions of the companies if they feel that their posts or accounts have been removed unjustly.



Facebook has taken this a step further with the creation of the Oversight Board, an independent entity which hears appeals decisions for users whose content or accounts have been taken off Facebook or Instagram.

The Oversight Board began accepting appeals in October 2020, and has since received 524,000 appeals cases. 36% of the appeals were for violations of Facebook's policy regarding hate speech, and another 31% were violations of bullying and harassment policies. Since its creation, the Oversight Board has decided 11 cases, overturning Facebook’s decisions all but 3 times and resulting in 52 recommendations to Facebook regarding policy changes to improve transparency in internal enforcement processes, as well as 35 instances of restored content.

Community Guidelines and Enforcement Methods

Furthering the goal of transparency, platforms publish Community Guidelines— unique sets of rules users agree to before using a site—as well as their procedures for enforcement.

In order to approach this successfully, all the companies included here publish periodic transparency reports to provide users with data about the type of content being taken down. Platforms also include appeals processes in their sites to allow for further review of violative content, empowering users to challenge the decisions of the companies if they feel that their posts or accounts have been removed unjustly.

Clearly articulated rules and procedures equip users with information needed to use social media effectively and hold platforms accountable for moderation decisions.

Twitter

Twitter's Rules include guidelines prohibit content involving:

- Safety

Such as violence, terrorism/violent extremism, child sexual exploitation, abuse/harassment, hateful conduct, suicide/self-harm, sensitive media, including graphic violence and adult content, illegal or certain regulated goods or services.

- Privacy

Such as private information of others to be shared without express permission, and does not allow non-consensual nudity.

- Authenticity

Such as spam or other forms of platform manipulation, interference in elections or other civic processes, impersonation of others, violations of intellectual property rights, or synthetic/manipulated media.

When a Tweet violates Twitter's Rules, enforcement options include:

- Labelling the Tweet as misinformation
- Limiting a Tweet's visibility in search results
- Requiring removal of the Tweet by the account
- Hiding the Tweet from public view.

If a particular account is continuously violating Twitter's Rules, enforcement options include:

- Requiring edits to the profile
- Placing accounts on read-only mode
  - So that the owner of the violating account cannot engage with other content
- Permanent suspension.

Facebook

Facebook's Community Standards prohibit content in the following categories:

- Violence and Criminal Behavior:

Including violence and incitement, dangerous individuals/organizations, publicizing crime, regulated goods, or fraudulent activity.

- Safety

Including content related to suicide and self-injury, child sexual exploitation, sexual exploitation of adults, harassment, human exploitation, or violations of personal privacy.

- Objectionable Content

Including hate speech, graphic violence, sexual content, and sexual solicitation.

- Integrity and Authenticity

Including false identities, spam, threats to cybersecurity, false news, manipulated audio or video material, and violations of intellectual property.

Facebook detects violations through a combination of artificial intelligence and human reviewers. When a post violates Facebook's Standards, enforcement options include:

- Warnings or disclaimers on sensitive or misleading content.
- Removing content and placing a strike on the account.
- Strikes remain on an account for 90 days for most violations, but can be held for up to 4 years in severe cases, such as violations of the dangerous individuals/organizations policy. Though severe violations may be subject to more severe action, strikes are commonly enforced as follows:
  - **1 strike:** Warning
  - **2 strikes:** One-day restriction from posting or commenting
  - **3 strikes:** 3-day restriction
  - **4 strikes:** 7-day restriction
  - **5+ strikes:** 30-day restriction
- Disabling accounts: Facebook will disable accounts with severe strikes or multiple rule violations.

Instagram

Instagram's Community Standards share many of the policies outlined by Facebook.

Instagram prohibits violations of intellectual property rights and copyright, and does not allow images involving nudity, compromising images of children, spam, fake reviews and ratings, impersonation of others, terrorism and organized crime, regulated goods, sexual services, adult sexual exploitation, hate speech, violence and incitement, harassment and bullying, content which promotes suicide and self-injury.

TikTok

TikTok's Community Guidelines outline types of content prohibited on the platform.

TikTok identifies prohibited content as that which involves violent extremism, hateful behavior, illegal activities, graphic violence, suicide or self-harm, harassment, nudity and sexual activity, any threat to or exploitation of minors, spam or fake accounts, or any activity that threatens the security of the TikTok platform.

TikTok removes content that violates these guidelines, notifying the user responsible for the post and providing an appeal process for those who believe their content was removed incorrectly.

- Severe or repeated violations result in account removal.

YouTube

YouTube's Community Guidelines prohibit content in the following categories, with additional guidelines for those who use their monetization feature to profit on content on the platform:

- Spam & Deceptive Practices

Including fake engagement, impersonation of others, external links to content that violates guidelines, spam and misinformation.

- Sensitive Content

Including content which endangers the emotional and physical well-being of minors, sexual content, promotion of suicide or self-injury or excessive vulgar language.

- Violent Or Dangerous Content

Including hate speech, harassment, graphic violence, violent criminal organizations.

- Regulated Goods

Including content relating to firearms or regulated goods and services.

YouTube uses artificial intelligence to flag content that violates their Community Guidelines, which is then looked at by human reviewers.

- YouTube channels may be removed for severe policy violations, though most violations will result in a warning, then a strike against the channel which prevents uploads for 1 week.
- Channels with 3 strikes within 90 days are removed.

Pinterest

Pinterest's Community Guidelines include provisions regarding content safety, intellectual property and spam.

Their content safety guidelines include prohibition of adult content, exploitation, hate speech, misinformation, harassment and criticism, private information, self-injury and harmful behavior, graphic violence and threats, violent actors, dangerous goods and activities, deceptive products and practices, and impersonation of others.

Content that violates the Community Guidelines is either removed or limited in distribution on the platform.

Snap

Snapchat's Community Guidelines prohibit sexually explicit content, harassment or bullying, threats or graphic violence, impersonation, illegal content, terrorism, hate groups or hate speech.

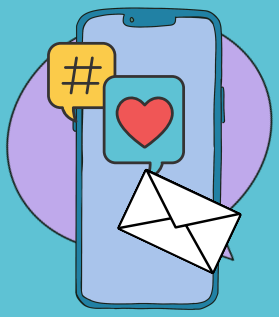
Content that violates these guidelines can be removed and could result in termination or limitations in the visibility of the account.

- Snapchat also notifies law enforcement in cases where laws have been violated.
- Once an account has been terminated, the individual is no longer allowed on the platform.

In Conclusion

Content moderation is an essential function of social media as users enjoy it today. Without the ability to take down posts that are illegal or inappropriate, using social media would not only be less enjoyable, but potentially far more dangerous.

The ease of posting on social media allows everyone to have a public voice, but those who abuse this voice to share harmful material such as graphic violence, terrorist material, or child sexual abuse highlight the need for platforms to have autonomy to remove content when necessary.



Social media companies have shown they understand their responsibility to users to maintain a safe online environment while also being transparent about procedures such that users feel confident in their ability to use their services.

The rate of violative content removed from platforms and the level at which it is removed prior to being seen by users makes clear companies are successfully prioritizing the safety of their users.