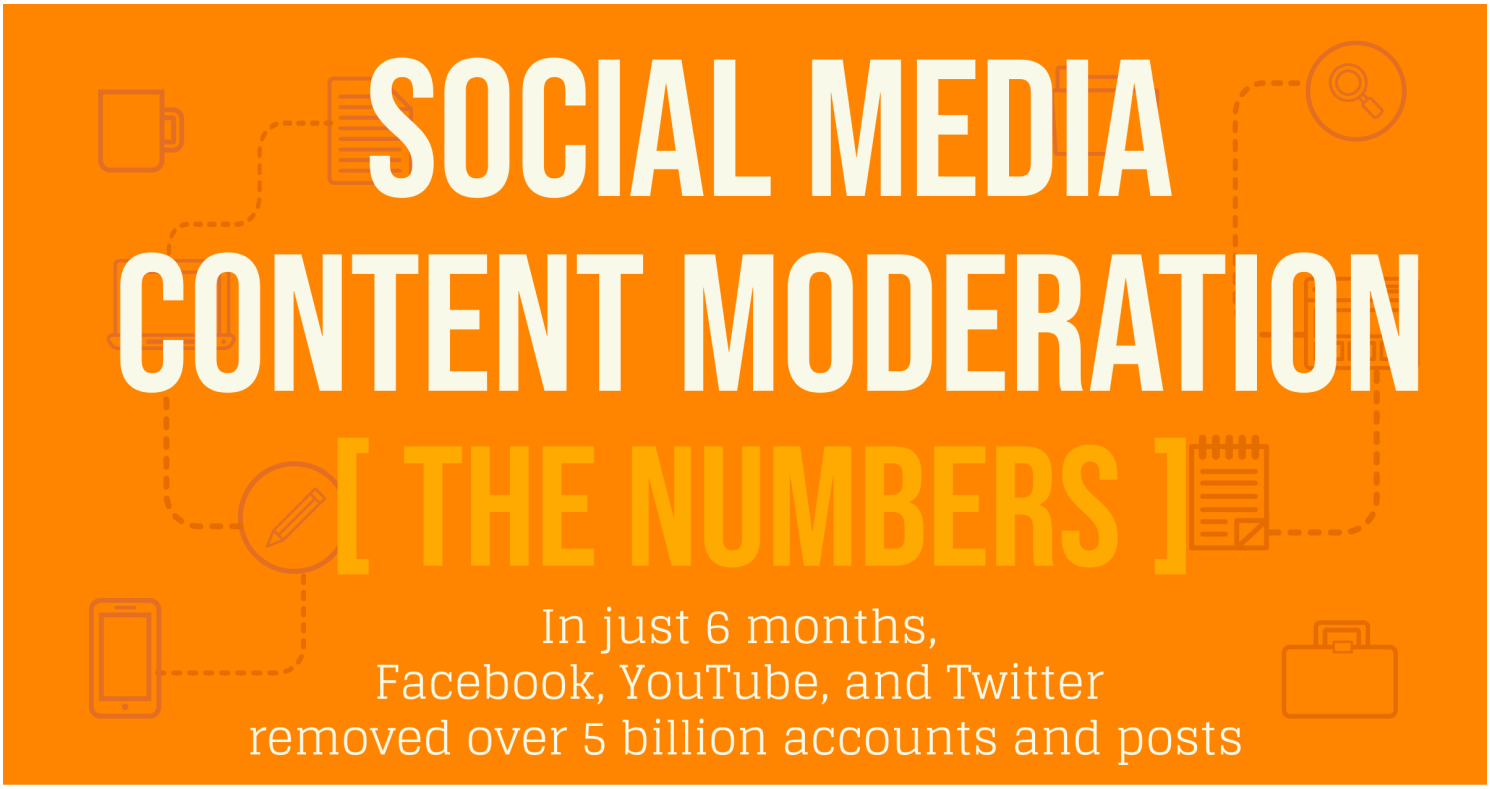


Social Media Companies' Transparency Reports on content moderation

This report aims to aggregate and clarify some of the findings and data from transparency reports by major social media platforms. As there is growing criticism of online platforms in Washington D.C., it's important that debate be as informed - including showing what content is removed from our largest social media platforms.

5 billion

Accounts and posts removed in just 6 months

“Despite what you may hear, platforms are actively removing offensive and objectionable content all the time”

- Carl Szabo, Vice President NetChoice

In just the six-months from July to December 2018, Facebook, Google, and Twitter took action on over 5 billion accounts and posts (5,051,079,936).

It broke down in the following ways:

- 17 million accounts and posts removed related to Child Safety (17,243,426)
- Over 57 million accounts and posts removed related to Pornography and Nudity (57,300,867)
- Nearly 2 billion accounts and posts removed related to Fake Accounts, Impersonations, and Doxxing (1,954,046,453)
- Over 3 billion accounts and posts removed related Spam (3,010,481,904)
- 12 million accounts and posts removed due to Extremist, Terrorist, and Hateful Conduct (12,007,286).
  The following report shows how much content the major social media platforms remove and why.

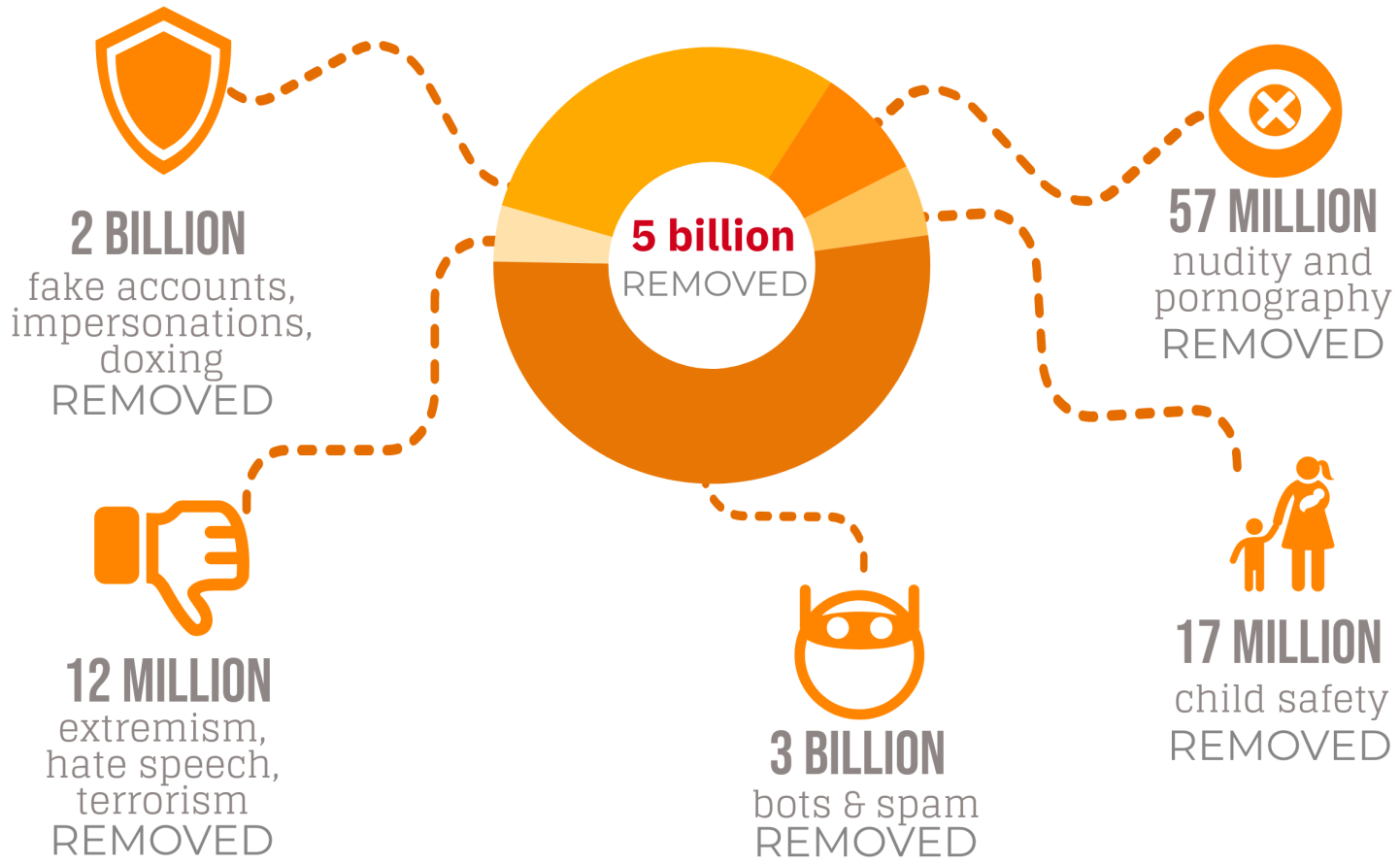
The report categorizes the reasons why content is removed, and compares the guidelines of different platforms. The report also shows how much content has been removed for each reason.

There are a number of categories in this list that feature legal speech, such as pornography, extremism, and impersonation. To remove such content, platforms relied on the Good Samaritan protections enabled bySection 230 of the Communications Decency Act. Unlike some forms of illegal content, Congress cannot mandate the removal of constitutionally protected speech.

SOCIAL MEDIA CONTENT MODERATION

[THE NUMBERS]

In just 6 months,
Facebook, YouTube, and Twitter
removed over 5 billion accounts and posts



2 BILLION
fake accounts, impersonations, doxing
REMOVED

12 MILLION
extremism, hate speech, terrorism
REMOVED

3 BILLION
bots & spam
REMOVED

57 MILLION
nudity and pornography
REMOVED

17 MILLION
child safety
REMOVED

Source: Twitter, Facebook, YouTube transparency reports (publicly available)

In this report, we will be including guidelines and takedown statistics from:

- Twitter
- YouTube
- Facebook

Their transparency reports can be found at the following links:

- https://transparency.twitter.com/en.html
- https://transparencyreport.google.com/youtube-policy/removals?hl=en&total_channels_removed=period:Y2018Q4&lu=total_channels_removed
- https://transparency.facebook.com/community-standards-enforcement

If you have any questions, please email info@netchoice.org or reach us on Twitter at @NetChoice.

Summary

The number of accounts banned, removed or otherwise actionned by YouTube, Twitter, and Facebook from July-December 2018 was 5,038,676,209.

The number of posts/comments/videos removed by YouTube, Twitter, and Facebook from October-December 2018 was 6,789,791.

Both of these statistics only include action taken because of guidelines covered in this report.

The most common reasons that accounts and posts were removed:

- Fake Accounts, Impersonation and Doxxing
- Spam
- Child Safety
- Porn and Nudity
- Extremist, Terrorist, and Hateful Conduct

Child Safety

A central role for all online platforms is to protect their most vulnerable users - children. Especially in an age where teens are often better at using online services than their parents, platforms play a crucial role in ensuring that young users are not manipulated or exploited online.

Similarly, platforms have an important role in ensuring that young people are not being exploited in the creation of online content - including everything from cyberbullying to child trafficking.

Accounts and posts removed by Facebook, Google, and Twitter: 15,953,903

Definitions:

Twitter:

"Twitter does not tolerate any material that features or promotes child sexual exploitation. This may include media, text, illustrated, or computer generated images." "Examples of content that depicts or promotes child sexual exploitation include, but are not limited to: (1) visual depiction of a minor engaging in sexually explicit or sexually suggestive act [sic]; (2) illustrated, computer-generated or other forms of realistic depictions of a human minor in a sexually explicit context, or engaged in a sexually explicit act; and (3) links to third-party sites that host child sexual exploitation material."

Google/YouTube:

"Content that endangers the emotional and physical well-being of minors is not allowed on YouTube." Examples: "(1) sexualization of minors; (2) harmful or dangerous acts involving minors; (3) infliction of emotional distress on minors; (4) misleading family content; and (5) cyberbullying and harassment involving minors."

Facebook:

"We do not allow content that sexually exploits or endangers children." "Do not post: (1) Content that depicts participation in or advocates for the suxual exploutation of children; (2) content that constitutes or facilitates inappropriate interactions with children; (3) content that depicts . . . sexual activity involving minors; (4) content that shows minors in a suxualized context; or (5) content that depicts child nudity."

Pornography and Nudity

Many tech businesses choose to make themselves appropriate and accessible for everyone by either banning or limiting the prevalence of pornography and nudity on their platforms, as all three in this report do.

Accounts and posts removed by Facebook, Google, and Twitter: 555,346,601

Definitions:

Twitter:

"Examples of content covered under this policy include: (1) graphic violence (e.g., media that depicts death or serious injury); (2) adult content (e.g., media that is pornographic or intended to cause sexual arousal); (3) intimate media (e.g., intimate photos or videos of someone distributed without their consent); and (4) hateful imagery (e.g., logos, symbols, or images whose purpose is to promote hostility and malice against others based on protected category)."

YouTube:

"Explicit content meant to be sexually gratifying (like pornography) is not allowed on YouTube. Videos containing fetish content will be removed or age-restricted. In most cases, violent, graphic, or humiliating fetishes are not allowed on YouTube."

Facebook:

"We restrict the display of nudity or sexual activity because some people in our community may be sensitive to this type of content."

Fake Accounts, Impersonation, and Doxxing

While often legal, impersonation and spreading someone's private information without permission can be very harmful to online users. Other than the more obvious risks, such as the loss of financial or personal security, impersonation can also undermine trust in online services.

Fake accounts often impersonate unconsenting people and threaten the security of a platform.

Accounts and posts removed by Facebook, Google, and Twitter: 1,954,041,182

Definitions:

Twitter:

"You may not publish or post other people's private information without their express authorization and permission. Definitions of private information may vary depending on local laws." Examples of private information include: (1) private identifiers or financial information, such as credit card information, social security or other national identity numbers; (2) locations of private residences or other places that are considered private; and (3) non-public personal contact information, such as phone numbers and email addresses.

YouTube:

"If someone has posted your personal information or uploaded a video of you without your consent, you can request removal of content based on our Privacy Guidelines." "For content to be considered for removal, an individual must be uniquely identifiable. If you want to use the privacy complaint process, make sure that you are uniquely identifiable within the content you seek to report before proceeding. When assessing if an individual is uniquely identifiable, we consider the following factors: (1) Image or voice; (2) Full name; (3) Financial information; (4) Contact information; and (5) Other personally identifiable information."

"Content intended to impersonate a person or channel is not allowed on YouTube. YouTube also enforces trademark holder rights. When a channel, or content in the channel, causes confusion about the source of goods and services advertised, it may not be allowed."

"Things like predatory behavior, stalking, threats, harassment, intimidation, invading privacy, revealing other people's personal information, and inciting others to commit violent acts or to violate the Terms of Use are taken very seriously. "

Facebook:

"Our goal is to remove as many fake accounts on Facebook as we can. We prioritize enforcement against users and accounts that seek to cause harm and find many of these fake accounts are used in spam campaigns and are financially motivated."

Extremism, Terrorism, and Hateful Content

Extreme political speech and everything up to explicit and direct threats of violence are constitutionally protected - but that doesn't mean we want to constantly encounter this sort of speech online. On many largely unmoderated platforms, that is increasingly the case. As a result, larger platforms remove speech that promotes white supremacy, terrorism, or hate of vulnerable groups.

Accounts and posts removed by Facebook, Google, and Twitter: 11,596,619

Definitions:

Twitter:

"You may not promote violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease." "We also do not allow accounts whose primary purpose is inciting harm towards others on the basis of these categories."

"[W]e do not allow users to make specific threats of violence or wish for the serious physical harm, death, or disease of an individual or group of people." "Examples of content covered under this policy include: (1) promoting terrorism; (2) soliciting bounties for serious violence; and (3) affiliating with and promoting organizations that use or promote violence against civilians to further their causes."

"Examples of content covered under this policy include: (1) graphic violence (e.g., media that depicts death or serious injury); (2) adult content (e.g., media that is pornographic or intended to cause sexual arousal); (3) intimate media (e.g., intimate photos or videos of someone distributed without their consent); and (4) hateful imagery (e.g., logos, symbols, or images whose purpose is to promote hostility and malice against others based on protected category)."

YouTube:

"Content or behavior intended to maliciously harass, threaten, or bully others is not allowed on YouTube."

"[W]e don't support content that promotes or condones violence against individuals or groups based on race or ethnic origin, religion, disability, gender, age, nationality, veteran status, or sexual orientation/gender identity, or whose primary purpose is inciting hatred on the basis of these core characteristics." "Examples: ""(1) Encourage violence against individuals or groups based on the attributes noted above. We don't allow threats on YouTube, and we treat implied calls for violence as real threats. You can learn more about our policies on threats and

harassment (2) Dehumanizing individuals or groups by calling them subhuman, comparing them to animals, insects, pests, disease, or any other non-human entity."""

"Violent or gory content intended to shock or disgust viewers, or content encouraging others to commit violent acts are not allowed on YouTube."

Facebook:

"Bullying and harassment happen in many places and come in many different forms, from making threats to releasing personally identifiable information, to sending threatening messages, and making unwanted malicious contact." Examples: repeated unwanted contact of a single person; repeated unsolicited contact of large numbers of people; malicious targeting; etc. "We define hate speech as a direct attack on people based on what we call protected characteristics — race, ethnicity, national origin, religious affiliation, sexual orientation, caste, sex, gender, gender identity, and serious disease or disability. We also provide some protections for immigration status." "We define attack as violent or dehumanizing speech, statements of inferiority, or calls for exclusion or segregation. We separate attacks into three tiers of severity."

Spam

Spam isn't just annoying, it can also be dangerous. Without smart moderation, spam can trick people into revealing personal information like phone numbers, bank details and social security numbers.

Accounts and posts removed by Facebook, Google, and Twitter: 3,008,027,695

Definitions:

YouTube:

"Spam, scams, and other deceptive practices that take advantage of the YouTube community aren't allowed on YouTube. We also don't allow content where the main purpose is to trick people into leaving YouTube for another site."

Facebook:

"Spam is a broad term to describe inauthentic content and behavior on Facebook that violates our Community Standards. It can be automated (published by bots or scripts) or coordinated (when an actor uses multiple accounts to spread deceptive content). Spammers aim to build audiences to inflate their content's distribution and reach, typically for financial gain."