European Commission

Information provided by the IT companies about measures taken to counter hate speech, including their actions to automatically detect content

November 2022

Directorate-General for Justice and Consumers

YouTube's response to hate speech

Hate speech is not allowed on YouTube and we are bringing significant attention to detection and removal of hateful content on our platform. Our hate speech Community Guidelines specifically prohibit content that encourages or glorifies violence against individuals or groups, or whose primary purpose is to incite hate against individual or group based on attributes including age, ethnicity, disability, gender, nationality, race, immigration status, religion, sex, sexual orientation, and veteran status.

We rely on a combination of humans and machines to detect hate speech content at scale. When we become aware of hate speech on our platform, we remove it. Between April and June 2022 we removed over 4.4 million videos for violating all of our Community Guidelines. Of those over 145k videos, more than 32k channels and over 64.5 million comments were removed for violating our hate speech policies.

Between January and March of this year, of the more than 3.8M videos we removed globally for violating our Community Guidelines, more than 15k videos that were uploaded from an IP address in the EU were in violation of our hate policies.

We also have a robust appeals process. When a YouTube creator's video is removed due to a policy violation, we provide a link with simple steps to appeal the decision. If a creator chooses to submit an appeal, it will be reviewed by a different member of our Trust and Safety team than made the original decision. We publish information about appeals and content reinstatements in our Community Guidelines Enforcement report (available here).

For more information about how YouTube tackles hate speech see our Community Guidelines, How YouTube Works and our Community Guidelines Enforcement Report.

TikTok response to hate speech

As a platform, we value being a signatory to the European Commission's Code of Conduct on Countering Illegal Hate Speech and the close cross-industry collaboration it fosters as we all work toward the common goal of eliminating hate online.

Exercises like this help deepen our cooperation with experts around Europe, and our teams are already hard at work to implement the lessons we have learnt as part of our ongoing efforts to strengthen our policies and enforcement strategies against hate.

These tests are just one of the ways in which we seek to be transparent about how we keep our platform safe for our global community. For example, our Community Guidelines Enforcement Report shares information about the volume and nature of content we remove for violating our Community Guidelines or Terms of Service. In the second quarter of 2022, of the 113 million videos we moderated globally, 1.7% were removed for violating our hateful behaviour policies

Our approach to tackling hate speech

We do not tolerate hate on TikTok, and we remove content that contains and promotes hate speech or involves hateful behaviour, as explained in our Community Guidelines. We use a combination of technology and human moderators to identify, review and take action against violative content or accounts.

Hateful behaviour is complex and ever-evolving, and we continually look for ways in which we can improve. This includes providing regular training for our safety professionals to help them better detect hateful behaviour, symbols, terms, and offensive stereotypes, and to help them identify and protect counter speech.

Partners are critical to our progress, and we consult academics and experts from across the globe to keep abreast of evolving trends and to help us regularly evaluate our approach.

We also partner with external organisations as we aim to harness the power of our platform to educate our community. For example, to mark Holocaust Memorial Day this year, we partnered with the World Jewish Congress and UNESCO to provide our global community with easy access to educational resources all year round, so they can learn more about the Holocaust and the Jewish Community.

TikTok thrives because of the diversity of our community, and every day we strive to provide a safe space where people feel welcomed and empowered to express themselves.

Overview of Twitter safety measures to tackle hate speech

The Twitter Rules exist to help ensure that all people can participate in the public conversation freely and safely, and include specific policies that explain the types of content and behaviour that are prohibited.

We believe we have a responsibility to the public – particularly during periods of crisis, such as the war in Ukraine – to proactively enforce our rules, preserve access to Twitter, elevate credible and reliable information, protect the privacy and safety of the people who use our service and others, and guard against efforts to manipulate the public conversation.

Abuse and harassment

Under our Abusive Behaviour policy, we prohibit content that harasses or intimidates, or is otherwise intended to shame or degrade others. We took action on 940,679 accounts during this reporting period. This is a 10% decrease from our last report and is in line with a 11% decrease in accounts reported under this policy during this period.

July-December 2021 data:

- Accounts actioned: 940,679 (number of unique accounts that were suspended or had some content removed for violating the Twitter Rules)
- Accounts suspended (number of unique accounts that were suspended for violating the Twitter Rules): 82,971
- Content removed (number of unique pieces of content (Tweets, profile image, banner, or bio) that Twitter required account owners to remove for violating the Twitter Rules): 1,344,061

Hateful conduct

We expanded our Hateful Conduct policy in December 2021 to prohibit dehumanizing speech on the basis of gender, gender identity and sexual orientation. During this period 104,565 accounts were suspended under this policy, representing a 22% decrease in account suspensions since our last report.

July-December 2021 data:

- Accounts actioned: 902,169
- Accounts suspended: 104,565
- Content removed: 1,293,178

All our data disclosures can be found in the Twitter Transparency Center.

Recent updates

- The new reporting flow with a symptoms-first approach makes it easier for people around the world to report unhealthy and unwanted content on Twitter.
- "Unmention" lets you untag yourself from a conversation, so you can no longer be mentioned in replies.
- Our Hateful conduct dehumanization policy now covers all protected categories including gender, gender identity, or sexual orientation.
- We've expanded the Private Information and Media policy to include 'private media' that is non-consensually shared with the intent to harass.

Facebook and Instagram's efforts against hate speech

We use a combination of expert content review teams and technology to detect and review hate speech at scale. When we're made aware of violating content we take action. We invest billions of dollars each year in these teams and technology to help keep our platforms safe. We have quadrupled - to more than 40,000 - the people working on safety and security. We've also pioneered the use of artificial intelligence technology to remove hateful content proactively, before users report it to us via our reporting tools. As well as participating in the European Commission's Code of Conduct on Countering Illegal Hate Speech, we publish quarterly Community Standards Enforcement Reports (CSER) to more effectively track our progress on how we enforce our policies, including those against hate speech and organized hate. As well as content removed, we publish data on the prevalence of hate speech on Facebook and Instagram. We believe prevalence is the most important metric when it comes to measuring our integrity efforts, as it captures not what we caught, but what we missed and what people actually saw. These quarterly reports show a tremendous improvement over the years in our ability to tackle hate.

On Facebook between April and June 2022:

• The prevalence of hate speech on Facebook - which is an estimate of the amount of hate speech people actually see on our service - was 0.02%. In other words, out of every 10,000 views of content on Facebook, just 2 included hate speech. This is down by around 80% since we first started to report prevalence in 2020, when prevalence was 0.1-0.11% on Facebook.

• We took action on 13.5 million pieces of hate speech content, of which 95.6% was proactively detected before it was reported. This is up from 1.6 million pieces of content we removed when we first began reporting these metrics in 2017 - of which 23.6% was found before a user reported it.

• 2.7 million pieces of content that we actioned for hate speech were appealed. 238k of those pieces of content were restored as a result.

On Instagram between April and June 2022:

• The prevalence of hate speech on Instagram was 0.01-0.02%. In other words, out of every 10,000 content views on Instagram, between 1 and 2 were hate speech.

• We took action on 3.8 million pieces of hate speech content, of which 91.2% was proactively detected before it was reported. This is up from 645k pieces of content we removed when we first began reporting this metric for Instagram in 2019 - of which 44.1% was found before a user reported it.

• 397k pieces of content that we actioned for hate speech were appealed. 47.8k of those pieces of content were restored as a result.

We are constantly looking at ways we can improve our approach to combating hate on our services, including refinements to our policies, our enforcement and the tools we give to users. To better address hate speech, we're deploying new AI systems and have also begun using warning screens to educate and discourage people from posting something that may include hate speech or bullying and harassment from being posted in the first place. Abuse of our products isn't static — and neither is the way we approach our integrity work. We know we're never going to be perfect in catching every piece of harmful content. But we're always working to improve, share more meaningful data and continue to ground our decisions in research.

Snapchat policy against hate speech

From the start, Snapchat was designed to be different. Snapchat was launched as a visual messaging app to provide an alternative to traditional social media. Snapchat founders wanted to build something different to capture the spontaneity and fun of real-life conversations among friends. That's why Snapchat opens directly to a camera, instead of a feed of content, and is focused on connecting people who are already friends in real life.

Snapchat was designed as a closed and curated platform—and the app was built in a way that limits the spread of hate speech, racism, and disinformation. The ephemeral-by-default nature of content on the platform is an important built-in safety feature that limits how broadly content can be spread. Additionally, unvetted content cannot 'go viral' as the ability for unmoderated content to reach a large audience is limited.

Clear and straightforward policy on hate speech

The Community Guidelines set the foundation for engaging on Snapchat. These guidelines clearly and explicitly prohibit hate speech or content that demeans, defames, or promotes discrimination or violence, as well as the use of our platform by terrorist, extremist and hate groups.

The policy is implemented in a straightforward and consistent manner throughout the platform.

Easily accessible and available in-app reporting tools for our users

Snapchat users experiencing or witnessing potential violations of the platform's policies are encouraged to report it right away. There are easily accessible in-app reporting tools where people can report specific Snaps and accounts. Reports are reviewed and actioned by Snap's Trust and Safety Team, which operates around the clock, 24/7.

Snapchat's latest transparency report notes that, during H2 2021, the platform enforced against 63,767 unique accounts (country data available here). We saw a 16% reduction in total hate speech reports and a 31% reduction in unique account enforcements for hate speech from H1 2021, showcasing a downtrend of existing hateful content on our platform. The average turnaround time for our Trust and Safety team to action a hate speech report was 12 minutes (down from 40 minutes for the same period last year).

Partnerships and external collaborations

We partner with best-in-class safety experts, researchers, threat analysts, NGOs and other civil society collaborators to help develop and improve our safety policies and materials.

In particular, since 2018, Snap's Safety Advisory Board (SAB) - a group of traditional online safety-focused non-profits and related organizations, technologists, academics, researchers, and survivors of online harms - has been providing critical feedback on fostering the safety and well-being of our Snapchat community. Thanks to the expert advice and guidance of our SAB members and their partnership, we've made progress over the last four years, implementing important awareness-raising and educational efforts.

Snap recently expanded its Safety Advisory Board (SAB) to include a wider diversity of geographies, safety related disciplines and areas of expertise. The 18 recently appointed members are experts in combating significant online safety risks and have broad experience across a range of safety-related disciplines.

Safety measures to counter hate speech content on Jeuxvideo.com (JV)

JV is a French media dedicated to video games news with thousands of forums that users can use to comment and contribute on various topics. As such, JV is fully committed to ensure a safe place for its community where everyone can freely communicate.

As such, the forum charter expressly prohibits hateful content, including without limitation any content that incites hatred, violence or discrimination on account of people's origin, race, religion, disability, gender, sexual orientation.

Since JV became a signatory to the European Commission's Code of Conduct on Countering Illegal Hate Speech in 2019, JV implemented various measures to strengthen its capacity to counter hate speech content by relying on a combination of human and technical measures.

When JV becomes aware of an illicit content on its website, JV removes it in an expeditious manner: between January 2022 and October 2022, 99% of the reported contents have been processed in less than 24 hours.

Since JV released a transparency report about its moderation, 16% of the reported contents have been related to hateful content and removed from the website. Users are also able to appeal all decisions made by JV and all these appeals are reviewed to ensure a fair processing.

The results of the 5th and 6th evaluation of the European Commission's Code of Conduct on Countering Illegal Hate Speech showed that (i) JV assessed 100% of notifications within 24 hours, (ii) removed 100% of hate speech contents, and (iii) responded with a feedback to 100% of the notifications received.

These positive results demonstrate the commitments of JV to counter hate speech content online and the willingness to ensure a safe environment for its community.