European Commission

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

7 October 2021

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 January and March 2021 we removed over 85k videos and over 43.6 million comments for violating our hate speech policies on YouTube. In the same quarter, approximately 76% of the videos uploaded that were removed for violating our Hate Speech policy were taken down before they had 10 views. During the same reporting period, of the 9.5M videos we removed for violating our Community Guidelines, more than 15k videos that were uploaded from an IP address in Europe were in violation of our hate policies.

We also have a robust appeals process. Users are able to appeal our decisions if they disagree with us. These appeals are re-reviewed and either upheld or the content is reinstated. Users receive a notification of the final outcome.

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

Countering hate speech on TikTok

Being transparent about how we keep our platform safe helps build trust and understanding with our community. We are pleased to share the results from our first evaluation as a signatory to the European Commission's Code of Conduct on Countering Illegal Hate Speech, which we joined in 2020.

While we welcome the positives from our performance, our teams are already hard at work to take the lessons we have learnt and implement changes to our systems and processes to help keep TikTok a safe space for our community.

In addition to publishing these results, 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 first quarter of 2021, of the around 62 million videos we removed globally, 2% were removed for violating our policies on hateful behaviour. We proactively removed 67% of hateful behaviour videos before they were even reported to us, and 84% were removed within 24 hours of being posted.

Our approach to tackling hate speech

Hate has no place on TikTok, as made clear in our Community Guidelines. We use a combination of technologies and moderation teams to identify and review content that would potentially violate these guidelines, and take actions including removing videos and comments and banning accounts.

We appreciate that hate speech is complex and ever-evolving, which is why we are constantly looking for ways in which we can improve. For example, last year, we strengthened our enforcement against hate speech to help ensure we capture the evolving landscape, language and terminology of hateful behaviours.

We also invest in regular training for our moderation teams to better detect hateful behaviour, symbols, terms, and offensive stereotypes and as a measure to help us properly identify and protect counter speech.

Partners are critical to our progress. Earlier this year, we established a European Safety Advisory Council to advise on our content moderation policies and practices, with representation from experts in discrimination, hate speech and violent extremist ideologies.

When it comes to safety, there is no finish line. We know there is always more we can and must do to improve our policies, processes and products to help keep TikTok a safe home for everyone, no matter who they are. Our goal is to eliminate hate, and we're committed to that goal for as long as it takes.

Overview of Twitter safety measures to tackle hate speech

Twitter's purpose is to serve the public conversation and protect the Open Internet.

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 behavior that are prohibited.

- Over 50% of our enforcement is now done proactively through machine learning without the people who use Twitter having to carry the burden of reporting.
- We continue to invest in new product capabilities that allow us to broaden the spectrum of enforcement decisions we're able to make beyond the binary 'take down leave up', including prompts, labelling Tweets that are misleading, labelling accounts with context and de-amplifying and limiting engagement on certain Tweets.

Abuse and harassment: There was a 34% decrease in the number of accounts actioned for violations of our abuse policy from January - June 2020:

- Accounts actioned: 398.057 (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): 72,139
- 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): 609.253

Hateful conduct: In the past 12 months Twitter has updated our Hateful Conduct policy to cover new facets of our dehumanization guidance, prohibit targeting groups (not just individuals) when inciting fear or spreading fearful stereotypes about protected categories, prohibit behavior that incites discrimination and harassment against a protected category and we recently updated our policies related to the denial of violent events and abusive references to violent events where protected categories are the primary victims. From January - June 2020, we saw a 35% decrease in the number of accounts actioned under our Hateful Conduct Policy.

- Accounts actioned: 635.415
- Accounts suspended: 127.954
- Content removed: 955.212

Facebook and Instagram's efforts against hate speech

Facebook and Instagram regularly publish Community Standards Enforcement Reports (CSER), providing metrics on how we enforce our policies, including those against hate speech and organized hate. Those quarterly reports show a tremendous improvement over the years in our ability to tackle hate. We've invested billions of dollars in people and technology to enforce our rules, and we have more than 35,000 people working on safety and security. Advancements in AI technologies have allowed us to take action on more hate speech from Facebook and Instagram over time, and find more of it before users report it to us. When we first began reporting our metrics for hate speech, in Q4 of 2017, our proactive detection rate was 23.6%. This means that of the hate speech we took action on, 23.6% of it was found before a user reported it to us. The rest was actioned after a user reported it. Today we proactively detect about 95% of hate speech content we take action on. Whether content is proactively detected or reported by users, we often use AI to take action on the straightforward cases and prioritize the more nuanced cases, where context needs to be considered, for our reviewers.

Overall, our latest hate speech enforcement data from Q2 2021 is:

On Facebook:

- We took action on 31.5 million pieces of hate speech content, compared to 25.2 million in Q1 2021. Of this content, 97.6% was proactively detected before it was reported;
- We took action on 6.2 million pieces of organized hate content, compared to 9.8 million in Q1 2021 (marking a return to pre-Q1 levels as we update our proactive detection technology). Of this content, 97.8% was proactively detected before it was reported;
- 1.4 million pieces of content we actioned for hate speech were appealed. 88.1k of those pieces of content were restored as a result.

On Instagram:

- We took action on 9.8 million pieces of hate speech content on Instagram, up from 6.3 million in Q1 2021. Of this content, 95.1% was proactively detected before it was reported;
- We took action on 367,000 pieces of organized hate content, up from 325,000 in Q1 2021. Of this content, 77.7% was proactively detected before it was reported;
- We did not publish data on appealed and restored content on Instagram in the latest CSER because, due to a temporary reduction in our review capacity as a result of COVID-19, we could not always offer people the option to appeal on Instagram. We still gave people the option to tell us they disagreed with our decision, which helped us improve our accuracy.

Last year, we also started sharing the prevalence of hate speech on Facebook for the first time. We believe that prevalence is one of the most useful metrics for understanding how often people see harmful content on our platform.

In Q2 2021, the prevalence of hate speech on Facebook decreased for the third quarter in a row, dropping from 0.05-0.06% to 0.05%. In other words, out of every 10,000 views of content on Facebook, 5 of them included hate speech. We evaluate the effectiveness of our enforcement by trying to keep the prevalence of hate speech on our platform as low as possible. Since we started reporting on this prevalence metric, we have registered a steady decrease.