Twitter Report: Staying safe and informed on Twitter during COVID-19

Table Of Contents:

Executive Summary

As the global community faces the COVID-19 pandemic together, Twitter is helping people find reliable information, connect with others, and follow what's happening in real time.

Throughout these unprecedented times, Twitter has continued to adapt and update our policies and enforcement, as well as increase transparency and share more data to ensure experts and the public can better analyse how discussion around COVID-19 continues to evolve. We have kept an updated blog with all relevant information on Twitter's efforts covid19.twitter.com and to date, **over 160 million people have visited the COVID-19 curated page, over two billion times.**

Below is an overview of the measures we have taken to protect the health of the public conversation while ensuring we are a collaborative and open partner in endeavours to address the challenging and changing online and offline issues society is facing.

- We launched an update tailored to serve the needs of the academic research community doing research with Twitter data, called the Academic Research product track. This will allow researchers to continue to study a huge range of issues including Numerous studies on COVID-19, Combating hate speech, and even Climate change.
- We expanded our COVID-19 misleading information policy to cover misleading information about vaccines, covering Tweets which advance harmful false or misleading narratives about COVID-19 vaccinations will be removed. We will label or place a warning on Tweets that advance unsubstantiated rumors, disputed claims, as well as incomplete or out-of-context information about vaccines.
- Our dedicated COVID-19 search prompt feature **has been expanded to over 80 countries worldwide, including 17 EU Member States, and is currently available in 29 languages**. This helps people who search for COVID-19 info find credible, authoritative content at the very top of their search page. We are in the process of updating these prompts so as to also include official information on COVID-19 vaccines, as it is already the case in Denmark and Spain.
- We currently have **273 prompts active in 99 countries worldwide,** including EU Member States, covering 12 issue areas. In January we introduced the prompt dedicated to gender-based violence in Finland.
- We have started to update the COVID-19 prompts to include specific information on COVID-19 vaccines in EU countries. We will continue this work in partnership with the relevant and willing public health organisations.
- In over 30 countries, we launched 'Twitter Events Pages' that bring together the latest Tweets from a number of authoritative and trustworthy government, media and civil society sources in local languages. We regularly update these pages to ensure that people are met with credible information on Twitter.
- On 12 January we published our latest transparency report that includes data from January 1, 2020, through June 30, 2020. During this reporting period, our teams took enforcement action against 4,658 accounts for violations of our COVID-19 misleading information policy. As we've further invested in technology, our automated systems challenged 4.5 million accounts that were targeting discussions around COVID-19 with spammy or manipulative behaviors.
- From 1 January 2021 to 31 January 2021 there were 864 Promoted Tweets that violated our

COVID-19 policy.

- In January we introduced Birdwatch, a pilot (currently in the US) of a new community-driven approach to help address misleading information on Twitter. It allows people to identify information in Tweets they believe is misleading and write notes that provide informative context. This approach has the potential to respond quickly when misleading information spreads.

This report contains information on policies, products, philanthropy activities and actions undertaken from 1 to 31 January 2021. Via the following links you can consult the reports submitted in July, August, September, October, November and December.

Helping people find reliable information

Initiatives to promote authoritative content and empower citizens

As countries all over Europe are facing a new surge of COVID-19 cases, we continue our efforts to protect the public conversation, elevate sources of reliable information, and build partnerships with governments and non-profit organisations.

In January 2020, before the official designation of the virus and in partnership with national public health agencies and the WHO, we launched a dedicated search prompt feature so that when somebody searches for COVID-19 they are met with credible, authoritative content at the very top of their search experience. We constantly monitor the conversation on the service to ensure that any keywords, including misspellings, generate the quality search results.

Prompts have been expanded to over 80 countries worldwide and available in 29 languages. In the EU, the prompt is active in: Austria, Belgium, Cyprus, Denmark, Estonia, Finland, France, Germany, Greece, Ireland, Italy, Latvia, Netherlands, Poland, Portugal, Spain, Sweden. In Belgium and Finland, a bilingual prompt was created.It is also available in the United Kingdom. *All countries in the EU were contacted and offered the opportunity to launch the prompt. This opportunity still stands.*

In 2020 the top hashtag used was #COVID19 which was **Tweeted nearly 400 million times.** This shows the value and importance of elevating credible information through search prompts. #Stayhome was the third biggest hashtag of the year, which highlights people sharing greater awareness and concerns for public safety, as well as promoting more ways to be active and entertained at home.

In over 30 countries, we launched 'Twitter Events Pages' that bring together the latest Tweets from a number of authoritative and trustworthy government, media and civil society sources in local languages.

We continued to elevate the conversation addressing safety and effectiveness of mask wearing with a series of Twitter Moments in English, Spanish and Portuguese and marketing campaigns and a customised emoji that can be activated with the hashtag #WearAMask, which was translated into 20 languages.

Vaccines

Fighting misinformation and disinformation around vaccines

As the world continues to fight the COVID-19 pandemic and the global distribution of vaccines is underway, people continue to turn to Twitter to discuss what's happening and find the latest authoritative public health information.

In previous reports, we shared our approach around the conversation surrounding COVID-19 on Twitter. In this report we provide information and guidance on our expanded approach to misleading information around COVID-19 vaccines. We prioritize the removal of the most harmful misleading information and we will label Tweets that contain potentially misleading information about the vaccines.

In the context of a global pandemic, vaccine misinformation presents a significant and growing public health challenge. We are focused on mitigating misleading information that presents the biggest potential harm to people's health and wellbeing. Twitter has an important role to play as a place for good faith public debate and discussion around these critical public health matters.

Under our COVID-19 misinformation policy, we already required the removal of Tweets that include false or misleading information about:

- The nature of the virus, such as how it spreads within communities;
- The efficacy and/or safety of preventative measures, treatments, or other precautions to mitigate or treat the disease;
- Official regulations, restrictions, or exemptions pertaining to health advisories; and
- The prevalence or risk of infection or death.

Following the expansion of this policy we will require people to remove Tweets which advance harmful false or misleading narratives about COVID-19 vaccinations, including:

- False claims that suggest immunizations and vaccines are used to intentionally cause harm to or control populations, including statements about vaccines that invoke a deliberate conspiracy;
- False claims which have been widely debunked about the adverse impacts or effects of receiving vaccinations; or
- False claims that COVID-19 is not real or not serious, and therefore that vaccinations are unnecessary.

Starting in early 2021, we will label or place a warning on Tweets that advance unsubstantiated rumors, disputed claims, as well as incomplete or out-of-context information about vaccines. Tweets that are labeled under this expanded guidance may link to authoritative public health information or the Twitter Rules to provide people with additional context and authoritative information about COVID-19.

We enforce this policy in close consultation with local, national and global public health authorities around the world, and will strive to be iterative and transparent in our approach. We remained focused

on helping people find credible health information, verifying public health experts, and updating our policies in an iterative and transparent approach.

Elevating authoritative information on vaccines

Twitter supports authoritative and credible information around the topic of vaccines - no matter whether authoritative and reliable sources will ultimately advocate for or against a vaccine. Our role is to ensure people have the credible information necessary to make informed decisions.

Today it is indeed more important than ever for people to be able to make informed decisions about their health and the health of their family. We understand the importance of vaccines in preventing illness and disease, and recognise the role that Twitter plays in surfacing credible public health information. As part of our efforts to protect the health of the public conversation, we are partnering with health organisations in Europe and worldwide to ensure that people seeking information about vaccinations on Twitter can easily access reliable and accurate information in their language.

A vaccine prompt, in partnership with national or federal public health agencies or (when not possible) the WHO, is currently available in 37 countries and 15 different languages, including Belgium, France Germany, Ireland, Norway, Spain, and the UK. The prompts direct people who search for keywords associated with vaccines to the webpage of the health organisation in charge, where authoritative and trustworthy information is provided. We are liaising with governments to roll out the prompt in additional EU countries.

To ensure we are surfacing credible public health information on the COVID-19 vaccine, we are working in partnership with EU Member States to update the COVID-19 prompts so as to include information on COVID-19 vaccines and the link to a dedicated official webpage. This option can be selected by the national or federal public health agency in every country and has already been implemented in Denmark and Spain.

In parallel, we continue to regularly update our curated COVID-19 pages to ensure people can continue to find accurate and up to date information around COVID-19 and vaccines from trustworthy and official sources.

Supporting government organisations on Twitter

In Spain we worked with the Ministry of Health to update the COVID-19 prompt so as to include information on COVID-19 vaccines and the National COVID-19 vaccination Strategy to ensure people

can continue to find accurate and up to date information around COVID-19 from trustworthy and official sources.

In Germany we hosted the second roundtable in our series "Twitter in Dialogue: Disinformation & the Vaccine Debate", which is a non-public, multi-stakeholder expert roundtable to discuss cross-sector, transdisciplinary challenges, observations and solutions regarding a constructive, empowering and fruitful discourse in the digital sphere. We reconvened with stakeholders from health institutions, academia and across government bodies to grasp the issues comprehensively. Twitter has been the driving force in getting these different stakeholders together for a proactive discussion around the challenges for public conversation in regard to COVID-19. In fact, our first roundtable on vaccines was hosted in October 2020 to help collect valuable information from different stakeholders early on to feed into our thinking, policies and priorities. The expert meeting in January was focused around possible cross-sector collaborations and projects as well as rapid coordination measures among the stakeholders of the group.

Addressing the wider impact of COVID-19

In addition to the numerous measures taken to promote reliable sources of information summarised above and detailed in the previous reports, we continue to engage in addressing new and emerging challenges that COVID-19 has exacerbated.

We currently **have 273 prompts active in 99 countries worldwide, including EU Member States, covering 12 issue areas.**

In January, we launched a prompt to fight gender-based violence in Finland. The prompt directs users who search for advice about gender-based violence to the hotline of the local partners organisation, where they can seek help and find the support they need. This prompt is also available in Belgium, Denmark, France, Germany, Ireland, Italy, Spain, and Sweden.

Holocaust Memorial Day

During the COVID-19 pandemic, several organisations denounced a rise in hateful and racist ideologies around the world and the spread of false conspiracy myths about minority groups. Around International Holocaust Remembrance Day, Twitter launched a customised emoji representing a candle, a symbol of light in the darkness, that could be activated with the hashtag #WeRemember and supported the World Jewish Congress's 'We Remember' campaign. To ensure that the lessons of the Holocaust are not forgotten, people were invited to write: "We Remember" on a sheet of paper, take a photo holding the sign and post in on social media. The campaign registered tremendous support from survivors, politicians, athletes, and the broader public.

The Ads for Good grant was also used to promote Virtual Commemoration from Auschwitz taking place on January 27th.

The top Promoted Tweet in the context of this pro-bono campaign gained 1,205,942 impressions, 288,578 media views, of which 55,595 organic and 232,983 promoted, and 39,180 engagements, of which 5,820 organic and 33,360 promoted.

Among the Tweets promoted in the context of this pro-bono campaign, the one above gained 461,064 impressions, of which 124,625 organic and 336,439 promoted. It also received 17,253 total engagements, of which 4,050 organic and 13,202 promoted, with media engagements leading the way (11,707 in total). The difference between organic and promoted data clearly shows the value of the Ads For Good.

In collaboration with the Holocaust Memorial Day Trust, Twitter launched a pro-bono promoted trend for the day using #LightTheDarkness. The campaign was widely supported on Twitter.

- The promoted trend resulted in 24M trend impressions and nearly 50K mentions of #LightTheDarkness.
- In addition, we were pleased to have Laura Marks, Chair of the Holocaust Memorial Day Trust to speak at an internal event hosted by the Twitter Business Resouce Group, @TwitterFaith.

We continued to support non-profit organisations on the platform to disseminate key public health and safety information around COVID-19, to promote media literacy, as well as to tackle issues that were exacerbated during the pandemic, such as mental health-related and discrimination issues.

Initiatives and tools to improve awareness

As the global community faces the COVID-19 pandemic together, Twitter is helping people find reliable information, connect with others, and follow what's happening in real time. In serving the public conversation, our goal is to make it easy to find credible information on Twitter and to limit the spread of

potentially harmful and misleading content. We are open about the challenges we are facing and the measures we're putting in place to serve the public conversation at this critical time.

Data Access

Twitter firmly believes in open data access to study, analyse, and contribute to the public conversation; which is why we continue to maintain a broad public API. Researchers use Twitter data to provide valuable feedback on how the online conversations and interactions evolve on and off Twitter. We continue to provide more accessible ways to make data and information publicly available to researchers.

Background: Since 2006, Twitter's APIs have given researchers and developers the opportunity to tap into what's happening in the world. Twitter's APIs are a unique data source for academics and are used around the world in a wide range of fields, from disaster management to political science, every day. Every major social science conference likely features multiple papers based wholly or largely on Twitter data. Our service is the largest source of real-time social media data, and we make this data available to the public for free through our public API. No other major service does this. You can find out more **here**.

Launch of the Academic Research product track on the new Twitter API

Since the Twitter API was first introduced in 2006, academic researchers have used data from the public conversation to study topics as diverse as the conversation on Twitter itself - from state-backed efforts to disrupt the public conversation to floods and climate change, from attitudes and perceptions about COVID-19 to efforts to promote healthy conversation online. Today, academic researchers are one of the largest groups of people using the Twitter API.

For over a decade, academic researchers have used Twitter data for discoveries and innovations that help make the world a better place. Over the past couple of years, we have taken iterative steps to improve the experience for researchers, like when we launched a webpage dedicated to Academic Research, and updated our Twitter Developer Policy to make it easier to validate or reproduce others' research using Twitter data. We have also made improvements to help academic researchers use Twitter data to advance their disciplines, answer urgent questions during crises, and even help us improve Twitter. An example is the launch in April 2020 of the COVID-19 stream endpoint, the first free, topic-based stream built solely for researchers to use data from the global conversation for the public good.

Over two years ago, we started our own extensive research to better understand the needs, constraints and challenges that researchers have when studying the public conversation. In October 2020, we tested this product track in a private beta program where we gathered additional feedback. This gave us a glimpse into some of the important work that the free Academic Research product track we're launching today can now enable.

> *"The Academic Research product track gives researchers a window into understanding the use of Twitter and social media at large, and is an important step by Twitter to support the scientific community."*
> *- Dr. Sarah Shugars, Assistant Professor at New York University*
> *"Twitter's enhancements for academic research have the potential to eliminate many of the bottlenecks that scholars confront in working with Twitter's API, and allow us to better evaluate the impact and origin of trends we discover."*
> *- Dr. David Lazer, Professor at Northeastern University*

As of 26 January, with the new Academic Research product track, qualified researchers have access to all v2 endpoints released to date, as well as:

- Free access to the full history of public conversation via the full-archive search endpoint, which was previously limited to paid premium or enterprise customers
- Higher levels of access to the Twitter developer platform for free, including a significantly higher monthly Tweet volume cap of 10 million (20x higher than what's available on the Standard product track today)
- More precise filtering capabilities across all v2 endpoints to limit data collection to what is relevant for your study and minimize data cleaning requirements
- New technical and methodological guides to maximize the success of your studies

The release of the Academic Research product track is just a starting point. This initial solution is intended to address the most requested, biggest challenges faced when conducting research on the platform. We are excited to enable even more research that can create a positive impact on the world, and on Twitter, in the future.

Further information in this blogpost and in this thread.

Press coverage:
ZDNet: Twitter ouvre son API à la recherche universitaire
Politico Morning Tech

Transparency

Meaningful transparency between companies, regulators, civil society, and the general public is fundamental to the work we do at Twitter — this transparency is a key tenet of our efforts to preserve and protect the Open Internet. In line with this philosophy, in August 2020 we launched our new Twitter Transparency Center to make our data easier to understand and analyse for those who access our biannual Twitter Transparency Report.

Our latest Twitter Transparency Report, launched on 12 January, includes data from January 1, 2020, through June 30, 2020.

The **COVID-19 pandemic** severely impacted business operations for all of us around the world. Given the changes in workflows, coupled with country specific COVID-19 restrictions, there was some significant and unpredictable disruption to our content moderation work and the way in which teams assess content and enforce our policies - a disruption that is reflected in some of the data presented today. We increased our use of machine learning and automation to take a wide range of actions on potentially abusive and misleading content, whilst continually focusing human review in areas where the likelihood of harm was the greatest.

In March, we launched a COVID-19 misleading information policy to further protect the health of the public conversation. Between January 1, 2020, through June 30, 2020 our teams took enforcement action against 4,658 accounts for violations of this policy. As we've further invested in technology, our automated systems challenged 4.5 million accounts that were targeting discussions around COVID-19 with spammy or manipulative behaviors.

We continued our zero-tolerance approach to **platform manipulation** and any other attempts to undermine the integrity of our service. During this latest reporting period, our teams saw a 54% increase in anti-spam challenges — an increase that is due in part to the proactive measures we put in place to protect the conversation around COVID-19. We also saw a 16% increase in the number of spam reports, compared to the last reporting period.

Twitter discloses state-backed actors' attempts to disrupt the conversation on the service. During this reporting period, we took action on more than 52,000 accounts that we reliably attributed to **information operations** originating within China, Russia, Turkey, Serbia, Honduras, Egypt, Indonesia, Ghana and Nigeria as well as a KSA-affiliated actor.

There was a 5% increase in the number of accounts removed for violations of our **terrorism and violent extremism** policies during this reporting period — 94% of those accounts were proactively identified. Our current methods of surfacing potentially violating content for human review include leveraging the shared industry hash database supported by the Global Internet Forum to Counter Terrorism (GIFCT).

We do not tolerate **child sexual exploitation** (CSE) on Twitter. CSE material is removed from the service without further notice and reported to The National Center for Missing & Exploited Children (NCMEC). As we have expanded our teams and increased operational capacity in this area, we saw a 68% increase in our enforcement under our Child Sexual Exploitation Policy.

In terms of Twitter Rules enforcement

- Targeted **harassment** of someone, or inciting other people to do so, is against the Twitter Rules. There was a 34% decrease in the number of accounts actioned for violations of our abuse policy.
- We saw a steady increase in the number of accounts actioned under our **Civic Integrity** Policy, as elections happened around the world during this reporting period. There was a 37% increase in the number of accounts actioned for violations of this policy during this reporting period.
- Over the six month reporting period and amidst the COVID-19 disruptions to workflow, we saw a 35% decrease in the number of accounts actioned under our **Hateful Conduct** Policy. In March 2020, our Hateful Conduct Policy expanded to cover new facets of our dehumanization guidance, specifically prohibiting language that dehumanizes people on the basis of age, disability, and disease.
- We do not permit people to promote, advocate, and persuade another individual to engage in **self-harm or suicide**. There was a 49% decrease in the number of accounts actioned for violations of our suicide or self-harm policy.
- We have clear rules around the sharing of **private information** on our service. During this reporting period, we continued to see an upward trend in our enforcement under this policy — up by 68%. This increase was due to our proactive efforts in this area.

Further information can be found here.

COVID-19 Guidance Enforcement

We are currently reviewing our processes and analysis to provide more accurate updates and data. We will update the following reports with updated information. Meanwhile, our Twitter Transparency Report is available here.

Birdwatch, a community-based approach to misinformation

In the context of our efforts to fight misinformation, we want to broaden the range of voices that are part of tackling this problem, and we believe a community-driven approach can help. That's why on 25 January we introduced Birdwatch, a pilot (currently in the US) of a new community-driven approach to help address misleading information on Twitter.

Birdwatch allows people to identify information in Tweets they believe is misleading and write notes that provide informative context. We believe this approach has the potential to respond quickly when misleading information spreads, adding context that people trust and find valuable. Eventually we aim to make notes visible directly on Tweets for the global Twitter audience, when there is consensus from a broad and diverse set of contributors.

In this first phase of the pilot, notes will only be visible on a separate Birdwatch site. On this site, pilot participants can also rate the helpfulness of notes added by other contributors. These notes are being intentionally kept separate from Twitter for now, while we build Birdwatch and gain confidence that it produces context people find helpful and appropriate. Additionally, notes will not have an effect on the way people see Tweets or our system recommendations.

To date, we have conducted more than 100 qualitative interviews with individuals across the political spectrum who use Twitter, and we received broad general support for Birdwatch. In particular, people valued notes being in the community's voice (rather than that of Twitter or a central authority) and appreciated that notes provided useful context to help them better understand and evaluate a Tweet (rather than focusing on labeling content as "true" or "false"). Our goal is to build Birdwatch in the open, and have it shaped by the Twitter community.
To that end, we're also taking significant steps to make Birdwatch transparent:

- All data contributed to Birdwatch will be publicly available and downloadable in TSV files
- As we develop algorithms that power Birdwatch — such as reputation and consensus systems — we aim to publish that code publicly in the Birdwatch Guide. The initial ranking system for Birdwatch is already available here.

We hope this will enable experts, researchers, and the public to analyze or audit Birdwatch, identifying opportunities or flaws that can help us more quickly build an effective community-driven solution.
We want to invite anyone to sign up and participate in this program, and know that the broader and more diverse the group, the better Birdwatch will be at effectively addressing misinformation. More details on how to apply here.

We know there are a number of challenges toward building a community-driven system like this — from making it resistant to manipulation attempts to ensuring it isn't dominated by a simple majority or biased based on its distribution of contributors. We'll be focused on these things throughout the pilot.

From embedding a member of the University of Chicago's Center for RISC on our team to hosting feedback sessions with experts in a variety of disciplines, we're also reaching beyond our virtual walls and integrating social science and academic perspectives into the development of Birdwatch.

We know this might be messy and have problems at times, but we believe this is a model worth trying. We invite you to learn alongside as we continue to explore different ways of addressing a common problem. Follow @Birdwatch for the latest updates and to provide feedback on how we are doing.

Press coverage:
Business Insider: Twitter's new 'Birdwatch' feature will let users to sift through and add notes to misleading tweets
Tech Crunch: Twitter's Birdwatch fights misinformation with community notes
The Guardian: Birdwatch: Twitter pilot will allow users to flag misinformation

Advertising on COVID-19

Twitter has restricted advertising containing implicit or explicit reference to COVID-19. More specifically, advertising containing implicit or explicit reference to COVID-19 is allowed when refers to adjustments to business practices and/or models in response to COVID-19 and support for customers and employees related to COVID-19, with the following restrictions:

- distasteful references to COVID-19 (or variations) are prohibited
- content may not be sensational or likely to incite panic
- prices of products related to COVID-19 may not be inflated
- the promotion of certain products related to COVID-19 may be prohibited
    - We currently prohibit the advertising of medical face masks and alcohol hand sanitisers (with or without reference to COVID). Other products may be added to this list and enforcement can be retroactive.
- the mention of vaccines, treatments and test kits is permitted, only in the form of information, from news publishers which have been exempted under the Political Ads Content policy.

Public Service Announcements related to COVID-19 from governments and supranational entities (for example, World Health Organisation), as well as trusted partners approved by the Public Policy team

are permitted. Also allowed are news related to COVID-19 from media publishers who have been exempted under the Political Ads Content policy.

For complete information about Twitter's Ads Policies, visit Twitter.com/adspolicy and Twitter Ads Policy update log.

Twitter released some guidelines on brand communication in times of crisis in order to help brands communicate with their customers, employees, and the broader ecosystem during the pandemic. The focus was on reflecting on what was appropriate in the tone and content of the communication.

Violations of COVID-19 advertising policy

Our advertising policy on COVID-19 is very strict so violations of that policy do not in any way imply misinformation or disinformation.

From 1 January 2021 to 31 January 2021 there were 864 Promoted Tweets that violated our COVID-19 policy.

- We reviewed a random sample of 100 ads from that cohort and did not find any which contained misinformation.
- We estimate that about 93% of the violating content during that time was detected by our automated systems, and approximately 7% was brought into human review and rejected for policy violations.

Useful links

Coronavirus: Staying safe and informed on Twitter

COVID-19: Our approach to misleading vaccine information

Insights from the 17th Twitter Transparency Report

Twitter Transparency Report