Twitter April update: Code of Practice on Disinformation

Twitter aims to provide greater transparency leading up to the European elections on our own efforts around election integrity and security, including how we are working with our partners, governments, industry peers, civil society, and researchers.

In each month preceding the European Parliament election, Twitter will provide a report outlining our efforts in that month. Not every report will focus on the same measures, but collectively they will provide a comprehensive view of our approach to the commitments we have made to ensuring the integrity of our services and enhancing access to healthy democratic discourse around the EU elections.

This report will elucidate upon our work throughout **April 2019** and provide insights into upcoming progress in the coming months.

Previous Twitter reports

- [Twitter Progress Report](#)
- [January 2019 update](#)
- [February 2019 update](#)
- [March 2019 update](#)

Developments in April
- **Enforcement of ad policies:** metrics on the enforcement of ads policies and EU political campaigning ads broken down by Member State.
- **Election guide for political campaigning advertisers:** our new guide for EU political advertisers.
- **Ads Transparency Center and Political Advertiser Certification:** an update on the ATC and the progress of certified political advertising ahead of the EU elections
- **Making it easier to access details on EU political campaigning ads:** details on new measures to improve transparency around political advertising
- **Development of Election integrity policy:** details about our new policy tackling the misuse of Twitter's services for the purpose of manipulating or interfering in elections

- **Dedicated reporting feature for misleading voting-related content:** details on our new dedicated reporting feature for content which aims to mislead users about voting related matters
- **Updates on health related measures:** an overview of our achievements in improving the health of the public conversation on the platform
- **Enforcements updates on addressing spam, malicious automation, and fake accounts:** figures from the latest Twitter Transparency Report
- **#WorldPressFreedomDay and #HerStory:** a new initiative to promote healthy and inclusive discourse on the platform
- **Supporting research into disinformation:** detailing our latest initiatives and events supporting fact checkers and researchers

Scrutiny of ad placements

Enforcement of ad policies

As noted in our March report, we have gathered details on the enforcement of our ads policies across EU Member States, highlighting the work our teams are doing to protect our users from misleading and potentially harmful advertisements, and ensuring brand safety for our advertisers, including EU political advertisers. We have provided detailed enforcement metrics, broken down by Member State in the areas focusing on:

- **Unacceptable business practices**: for example prohibiting advertising for accounts making misleading, false, or unsubstantiated claims
- **Quality ads policy enforcement**: our ongoing review of accounts to ensure that advertisements adhere to editorial guidelines and quality reviews
- **EU political ads enforcement:** ads prevented from targeting EU Member States from non-certified accounts

Unacceptable Business Practices Ads policy

Twitter prohibits the promotion of unacceptable business practices globally. Examples of unacceptable business practices include:

- Potentially deceptive, misleading, or harmful business propositions
- Making misleading, false, or unsubstantiated claims
- Promoting misleading information or omitting vital information on pricing, payment terms, or expenses the user will incur
- Promoting offers or deals that are not available or easily located on the landing page

The table below illustrates the number of ads targeting EU Member States which were rejected per Twitter's Unacceptable Business Practices Ads Policy in the period 01 January 2019 - 31 March 2019.

Targeted Country	Number of ads
Austria	142
Belgium	254
Bulgaria	87
Croatia	65
Czech Republic	92
Denmark	123
Estonia	51
Finland	145
France	395
Germany	361
Greece	147
Hungary	74
Iceland	58
Ireland	129
Italy	279
Latvia	63
Lithuania	60
Luxembourg	65
Malta	62
Netherlands	339
Poland	116
Portugal	136
Romania	90
Slovakia	60
Slovenia	65
Spain	235
Sweden	411
United Kingdom	486

Quality Ads policy

To ensure that users have a positive experience on Twitter, advertisements must adhere to editorial guidelines for the following categories:

- User bio and destination URL
- Content, clarity and accuracy
- Text, image and video

The table below indicates the number of ads prevented from targeting EU Member States, rejected per Twitter's Quality Ads Policy in the period 01 January 2019 - 31 March 2019.

Targeted Country	Number of ads
Austria	216
Belgium	289
Bulgaria	166
Croatia	161
Czech Republic	199
Denmark	207
Estonia	153
Finland	250
France	461
Germany	338
Greece	222
Hungary	166
Iceland	122
Ireland	225
Italy	340
Latvia	155
Lithuania	133
Luxembourg	171
Malta	152
Netherlands	322
Poland	211

Portugal	230
Romania	188
Slovakia	140
Slovenia	155
Spain	382
Sweden	358
United Kingdom	1,421

EU Political Campaigning Ads policy

Political Content includes political campaigning for the upcoming EU Parliamentary elections, and political campaigning advertisers must complete a certification process. More details on the process can be found here.

The table below illustrates the number of ads prevented from targeting EU Member States from non-certified accounts during the first month of policy enforcement, beginning from the date of enforcement 11 March 2019 - 11 April 2019.

Targeted Country	Number of ads
Belgium	2
France	3
Germany	1
Netherlands	2
Spain	1
Sweden	1
United Kingdom	2
Total	**12**

Elections Guide for Political Campaigning Advertisers

In line with our commitment to assist and/or allow political campaigning advertisers to assess media buying strategies and online reputational risks, we have produced a guide for EU political advertisers.

Our interactive Guide to Twitter Advertising provides the answers to the questions most frequently asked, and helps EU political advertisers create compelling Tweets for their audiences before, during and after the 2019 European Elections. The guide includes details of our new EU Political Campaigning Policy, how to get certified and started with Twitter Ads, plus hints and tips on how to target audiences and measure campaign performance.

Download the guide here.

Transparency of advertising

Ads Transparency Center and Political Advertiser Certification

The Ads Transparency Center (ATC) continues to be populated as more of the transnational EU political parties and groups undertake the political advertiser certification process and begin to roll out their political campaigns ahead of the EU election. To date 11 advertisers have certified, and six are running ads which can be viewed here.

Making it easier to access details on EU political campaigning ads

We have recently added new features to the EU political ads on Twitter to make it easier to direct users from the political campaigning ad directly to the EU political Ads Transparency Center. This helps to ensure greater transparency and real-time information around the content and messaging leading up to the EU elections.

When a user views an EU political ad, the ad will already be labelled as a promoted post with the additional context of being (political). By clicking on the add you will be able to view further details including who is paying for the add, and be provided with a link (Learn more) which will direct users to the Ads Transparency Center.

Integrity of Services

Development of Election integrity policy

In April, we established an election integrity policy aimed at halting the misuse of Twitter's services for the purpose of manipulating or interfering in elections. This includes posting or sharing content that may suppress voter turnout or mislead people about when, where, or how to vote.

The public conversation occurring on Twitter is never more important than during elections. Any attempts to undermine the integrity of our service is antithetical to our fundamental rights and undermines the core tenets of freedom of expression, the value upon which our company is based. We believe we have a responsibility to protect the integrity of those conversations from interference and manipulation. Therefore, we prohibit attempts to use our services to manipulate or disrupt elections, including through the distribution of false or misleading information about the electoral process or when or how to vote.

We prohibit three categories of manipulative behavior and content under this policy:

Misleading information about how to participate

You can't share false or misleading information about how to participate in an election or other civic event. This includes but is not limited to: misleading information about how to vote or register to vote (for example, that you can vote by Tweet, text message, email, or phone call); misleading information about requirements for voting, including identification requirements; and misleading statements or information about the official, announced date or time of an election.

Voter suppression and intimidation

You can't share false or misleading information intended to intimidate or dissuade voters from participating in an election. This includes but is not limited to: misleading claims that polling places are closed, that polling has ended, or other misleading information relating to votes not being counted; misleading claims about police or law enforcement activity related to polling places or elections; misleading claims about long lines or equipment problems at voting locations; misleading claims about voting procedures or techniques which could dissuade voters from participating in an election; and threats regarding voting locations (note that our violent threats policy may also be relevant for threats not covered by this policy).

False or misleading affiliation

You can't create fake accounts which misrepresent their affiliation, or share content that falsely represents its affiliation, to a candidate, elected official, political party, electoral authority, or government entity. Content that would not be in violation of this policy include inaccurate statements about an elected official, candidate, or political party; organic content that is polarizing, biased, hyperpartisan, or contains controversial viewpoints expressed about elections or politics; discussion of public polling information; and using Twitter pseudonymously or as a parody, commentary, or fan account to discuss elections or politics.

Dedicated reporting feature for misleading voting related content

Further to the development of the election integrity policy, we created a dedicated reporting feature for content that aims to mislead users about issues pertaining to voting. This allows users to more easily report this content to us. This is in addition to our existing proactive approach to tackling malicious automation and other forms of platform manipulation on the service. We started with **EU and Indian elections** and plan to roll out to other elections globally throughout the rest of the year.

This function was fully operational in the EU as of 29 April.

Content which would violate this policy includes but is not limited to:

- Misleading information about how to vote or register to vote (for example, that you can vote by Tweet, text message, email, or phone call);

- Misleading information about requirements for voting, including identification requirements; and

- Misleading statements or information about the official, announced date or time of an election.

Details on how to report content that aims to mislead users about issues pertaining to voting on desktop and on mobile below ([Tweet here](#)):

Twitter Safety ✔ @TwitterSafety · Apr 24

Public conversation on Twitter is never more important than during elections. Today, we're launching a new reporting feature to tackle deliberate attempts to mislead about voting. We'll start with #LokSabhaElections2019 🇮🇳 & #EUelections2019 🔔

blog.twitter.com/en_us/topics/c…
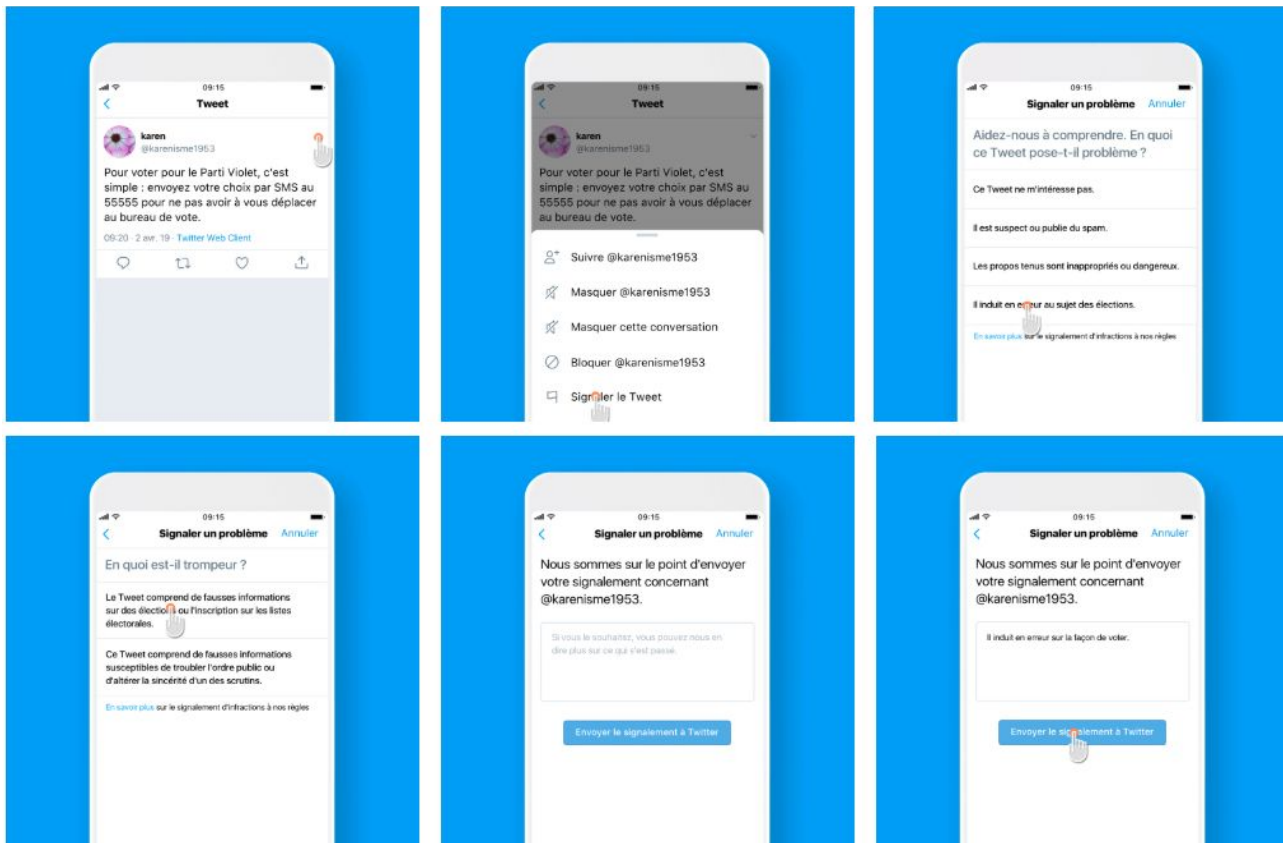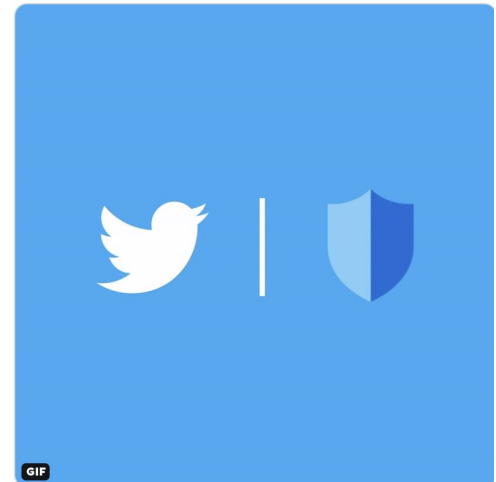
Updates on health-related measures

Last year, we shared that building a Twitter free of abuse, spam and other things that distract from the public conversation is our top priority. Since then, we've made strides in creating a healthier service. In April, we shared an update on our progress and previewed some changes that can be anticipated in the coming months. There will always be more to do, but we've made meaningful progress that is important to share. Here is a summary of progress we've made:

- 38% of abusive content that's enforced is surfaced proactively to our teams for review instead of relying on reports from people on Twitter

- 16% fewer abuse reports after an interaction from an account the reporter doesn't follow

- 100,000 accounts suspended for creating new accounts after a suspension during January-March 2019, a 45% increase from the same time last year

- 60% faster response to appeals requests with our new in-app appeal process

- 3 times more abusive accounts suspended within 24 hours after a report compared to the same time last year

- 2.5 times more private information removed with a new, easier reporting process

People who don't feel safe on Twitter shouldn't be burdened to report abuse to us. Previously, we only reviewed potentially abusive Tweets if they were reported to us. We know that's not acceptable, so earlier this year we made it a priority to take a proactive approach to abuse in addition to relying on people's reports.

This time last year, 0% of potentially abusive content was flagged to our teams for review proactively. Today, by using technology, **38% of abusive content that's enforced is surfaced proactively for human review instead of relying on reports from people using Twitter.** This encompasses a number of policies, such as abusive behavior, hateful conduct, encouraging self-harm, and threats, including those that may be violent.

The same technology we use to track spam, platform manipulation and other rule violations is helping us flag abusive Tweets to our team for review. With our focus on reviewing this type of content, we've also expanded our teams in key areas and geographies so we can stay ahead and work quickly to keep people safe. Reports give us valuable context and a strong signal that we should review content,

but we've needed to do more and though still early on, this work is showing promise.

Changes to come

Keeping people safe on Twitter remains our top priority, and we have more changes coming to help us work toward that goal. Here's some of what to expect in the coming months:

- We'll continue to improve our technology to help us review content that breaks our rules faster and before it's reported, specifically those who Tweet private information, threats, and other types of abuse.

- We'll make it easier for people who use Twitter to share specifics when reporting so we can take action faster, especially when it comes to protecting people's physical safety.

- Context on Tweets and our enforcement is important in understanding our rules, so we'll add more notices within Twitter for clarity, such as if a Tweet breaks our rules but remains on the service because the content is in the public interest.

- We are updating our rules in the next few weeks so they're shorter, simpler and easier to understand.

- Starting in June, we'll be experimenting with ways to give people more control over their conversations by giving them an option to hide replies to their Tweets.

Enforcements updates on addressing spam, malicious automation, and fake accounts

Platform manipulation refers to the use of Twitter to mislead others and/or disrupt their experience by engaging in bulk, aggressive, or deceptive activity. This activity includes, but is not limited to, spam, malicious automation (malicious use of bots), and inauthentic account abuse (fake accounts).

Further to the updates shared in our March report on the number of accounts challenged internally for spammy or malicious behaviour on the platform as well as the number of accounts reported by users, in April 2019 Twitter proactively challenged 13,787,154 accounts.

Proactive Anti spam* and fake account challenges

Month	Number of accounts challenged
January 2019	19,522,273
February 2019	17,036,751
March 2019	16,552,753
April 2019	13,787,154
January 2019 - April 2019	**66, 898, 931**

*Anti spam challenges refer to our internal, proactive detection of suspected spam and fake accounts. If we have reason to suspect an account may be engaging in platform manipulation, we may require the account holder to complete a challenge to confirm that a human is in control of the account. As stated above, this could be confirming a phone number or email address, or completing a reCAPTCHA challenge.

User reports of spam*

Month	Number of reports submitted
January 2019	489,148
February 2019	406,162
March 2019	504,729
April 2019	597, 295
January 01 - March 31 2019	**1,997,334**

*User reports of spam reflects reports submitted by people who use Twitter after receiving interaction from a suspected spam account. The lower numbers reflect our ongoing motivation to reduce the burden on users of Twitter to report spam and malicious automation.

To reiterate, it is important to note that this is raw data, these figures will be subject to change as it is reviewed for publication in our Transparency Report. We have provided it well ahead of time as we are committed to providing more transparency on our actions in the lead up to the EU elections.

Twitter Transparency Report (July - December 2018)

In April, we finalised the latest edition of the Twitter Transparency Report for the period of July - December 2018, which includes detail on platform manipulation. This refers to the use of Twitter to mislead others and/or disrupt their experience by engaging in bulk, aggressive, or deceptive activity. This activity includes, but is not limited to, spam, malicious automation (malicious use of bots), and inauthentic account abuse (fake accounts). These updates correspond to our biannual report and help to provide a larger picture of the work Twitter is carrying out to tackle platform manipulation on a daily basis, beyond the EU elections period.

Promoting the health of the public conversation by countering all forms of platform manipulation continues to be a top priority at Twitter. This report includes information about the number of anti-spam challenges issued and the number of reports of spam experienced by people on Twitter, broken down by month.

During this reporting period, we particularly focused on detecting fake accounts at sign-up. We also launched an enhanced spam reporting flow in October 2018 to provide people who use Twitter with the option to tell us more about the type of spam they experienced when they make a report. As manipulation and spam tactics are constantly evolving, this granularity in reports will help us refine our detection signals so that we can improve upon our ability to remove manipulative content before it affects the experience of people who use Twitter.

Number of accounts challenged monthly Jul 1 - Dec 31, 2018

Month	Challenges
July	37,508,396
August	30,474,937
September	37,257,274
October	40,671,724
November	25,489,160
December	22,185,461
TOTAL	193,586,952

Aggregate reports of spam have decreased in the second half of 2018, suggesting that people continue to experience fewer spammy interactions on Twitter. Despite the decrease in aggregate, reports received increased month over month from September through December. Reports of spam do not necessarily correspond with an increase in the overall volume of spam; one trend we observe at Twitter is that users overall tend to submit more reports, including spam, around major world events.

Spam challenges
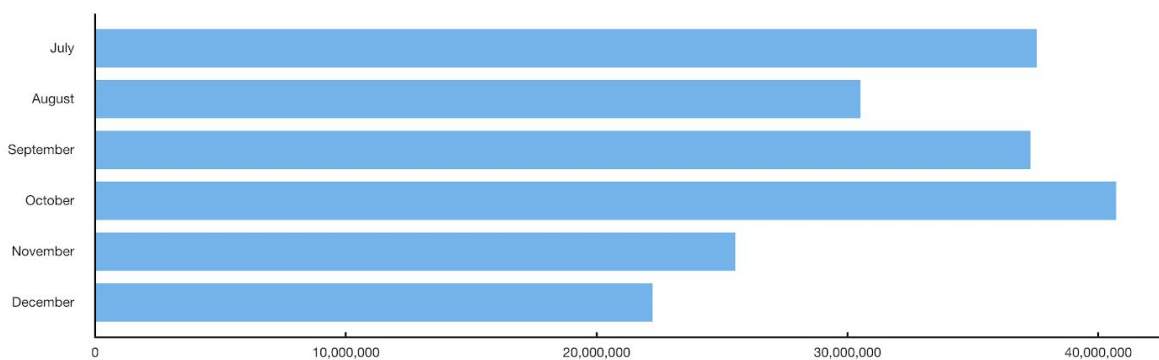
# Expanding digital literacy and empowering Consumers

#WorldPressFreedomDay

To celebrate #WorldPressFreedomDay 2019, we launched #HerStory, a monthly video series seeking to elevate female voices in journalism. #HerStory featured women journalists with diverse backgrounds and focus areas, who discussed about their lives, how they use Twitter to report the news, and other women journalists who inspire them. This initiative is part of our ongoing commitment to promote healthy and inclusive discourse on our platform.

Find out more about #HerStory here.

Twitter News ✔
@TwitterNews

We're excited to announce the launch of #HerStory, Twitter's new, monthly original video series highlighting the work and personal stories of women journalists around the world. Here's a sneak peak:

#HerStory
Twitter's new, monthly original video series

3M views   0:03 / 1:04
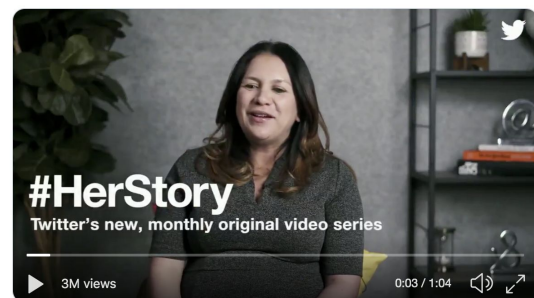
From **Twitter News** ✔

9:59 PM · May 2, 2019 · Twitter Web Client

236 Retweets   **1.6K** Likes

Empowering the research community

Supporting research into disinformation

Twitter España ✔
@TwitterEspana

Para finalizar la conversación de #ElPoderMiVoz, @cjimenezcruz, Fundadora de @maldita_es, pone luz a la desinformación en cuestiones de género: "Es muy importante conocer los datos de las cuestiones de género para poder confrontar a los comentarios machistas a día de hoy."

Translate Tweet

11:08 AM · Mar 22, 2019 · Twitter Web Client

We continue to support and promote disinformation research efforts, and to actively engage with the research community. In Spain, we hosted an event as part of our #PositionOfStrength / #ElPoderMiVoz initiative for women's empowerment, in participation with the fact checking organisation Maldita (members of the International Fact Checking Network and High-Level Expert Group on Fake News and online disinformation).

We also provided support for their project Maldito Feminismo, which focuses on debunking misinformation regarding women.