European Commission

Countering illegal hate speech online

7th evaluation of the Code of Conduct

Factsheet | November 2022

Didier Reynders
Commissioner for Justice

Directorate-General for Justice and Consumers

The seventh evaluation on the Code of Conduct on Countering Illegal Hate Speech Online shows that the number of notifications reviewed within 24 hours (64.4%) has decreased as compared to 2021 (81%) and 2020 (90.4%). Only TikTok has increased its performance (from 82.5% to 91.7%). The average removal rate (63.6%) is similar to 2021 (62.5%), but still lower than in 2020 (71%). Looking at the individual performance of the platforms, most of them (except for YouTube) have removed less hate speech content than in 2021. The quality of feedback to users' notifications has improved as compared to previous monitoring exercises.

Key figures

1. Notifications of illegal hate speech

- **36 organisations** from 21 Member States **sent notifications to the IT companies** taking part in the Code of Conduct regarding hate speech deemed illegal during the period 28 March to 13 May 2022.[1]
- A total of **3634 notifications** were submitted to the IT companies part of the Code of Conduct.
- **2765 notifications** were submitted through the **reporting channels available to general users**, while **869** were submitted through **specific channels available only to trusted flaggers/reporters**.
- **Facebook** received the largest amount of notifications **(1558)**, followed by Twitter **(1097)**, YouTube **(423)**, Instagram **(398)**, TikTok **(151)** and Jeuxvideo.com **(7)**.[2] Snapchat, Dailymotion and Microsoft did not receive any notification in the course of the monitoring exercise. LinkedIn, which joined the Code in 2021, will be monitored from the 2023 monitoring exercise.
- In addition to flagging the content to IT companies, the organisations taking part in the monitoring exercise sent **176 cases of hate speech to the police, public prosecutor's bodies or other national authorities.**

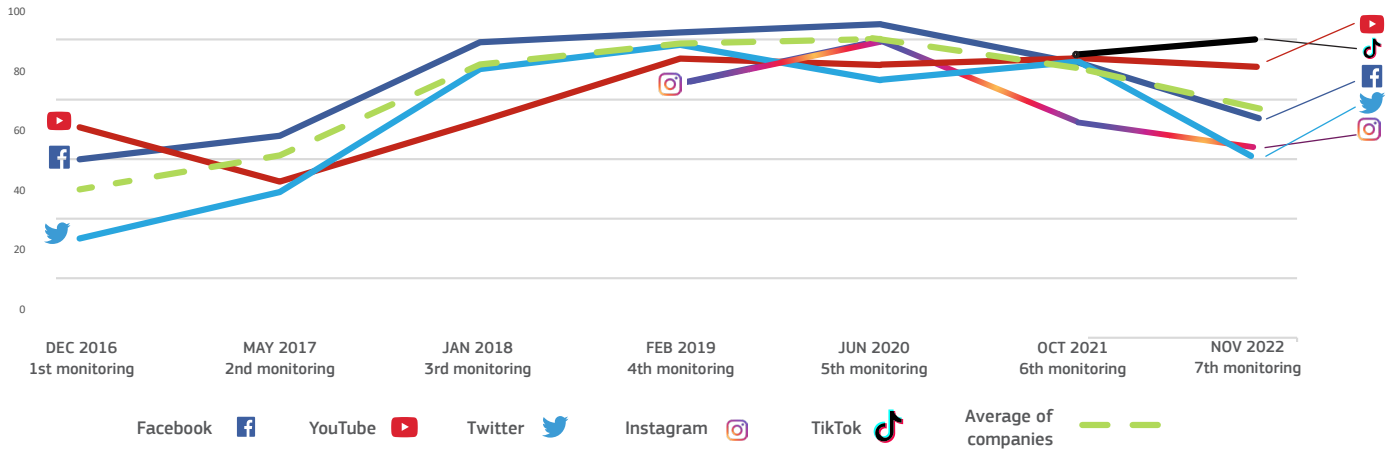
[1] In order to establish trends, this exercise used the same methodology as the previous monitoring rounds (see Annex).
[2] Given the very low amount of notifications, the performance of Jeuxvideo.com on removal rates, time of assessment and feedback to users is not reflected in this report. All the seven notifications sent to Jeuxvideo.com were assessed within 24h, the users received a feedback and content was removed.

2. Time of assessment of notifications

- In **64.4% of the cases,** the IT companies assessed the notifications **in less than 24 hours,** an additional **12.7%** in less than 48 hours, **21.5%** in less than a week and in **1.4%** of cases, it took more than a week.
- **The Code of conduct prescribes that the majority of notifications is assessed within 24 hours.** All IT companies are therefore on target, yet, the average results are lower than in 2021 and 2020 (**81% and 90.4%, respectively**).

TikTok assessed notifications in less than 24 hours in **91.7%** of the cases and an additional **3.8%** in less than 48 hours. The corresponding figures for YouTube are **83.3%** and **7%** and for Twitter **54.3%** and **28.9%**, respectively. Instagram had **56.9%** and **5.9%**, and Facebook **63.8%** and **8.2%**. Only TikTok had a better performance than in 2021, while all other platforms had a worse score than last year.

Percentage of notifications assessed within 24 hours - Trend over time

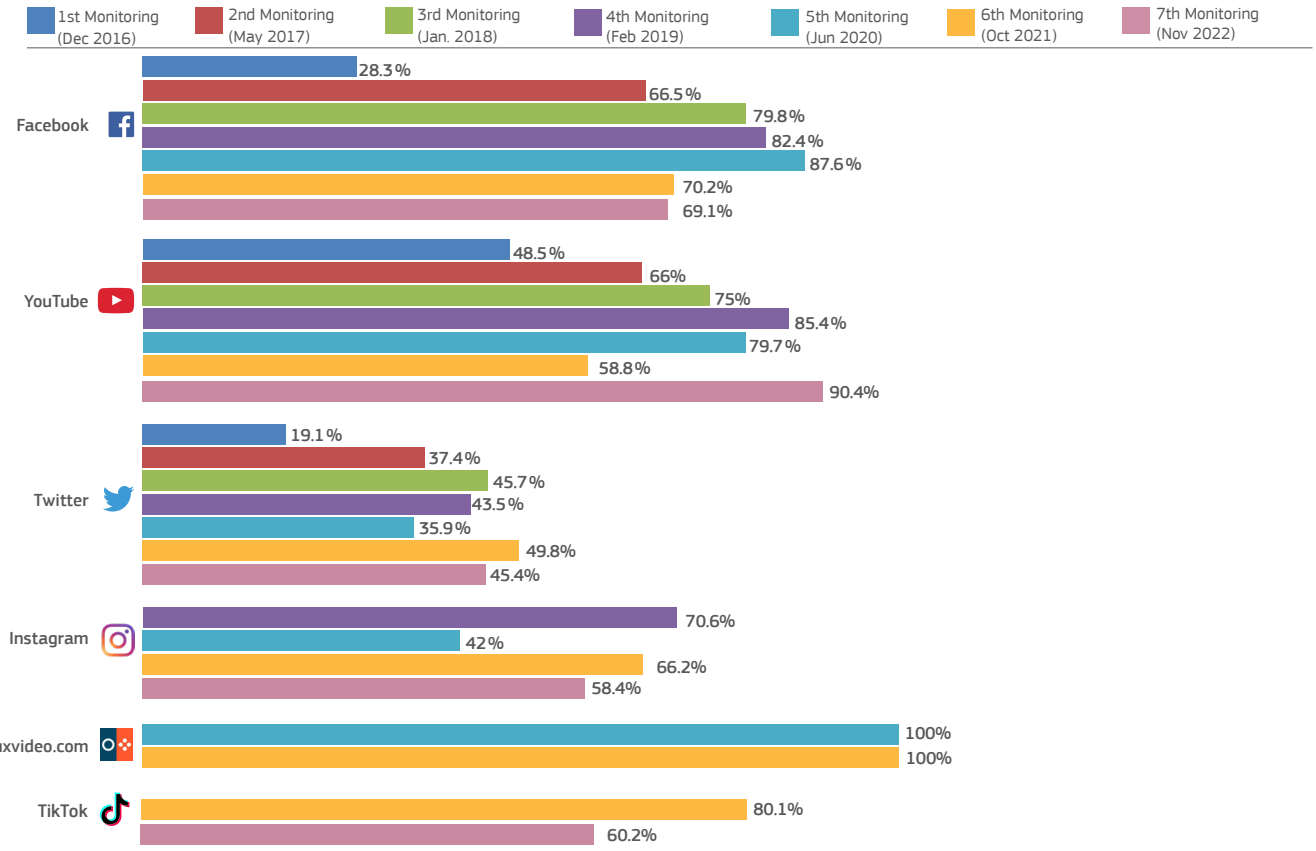


3. Removal rates

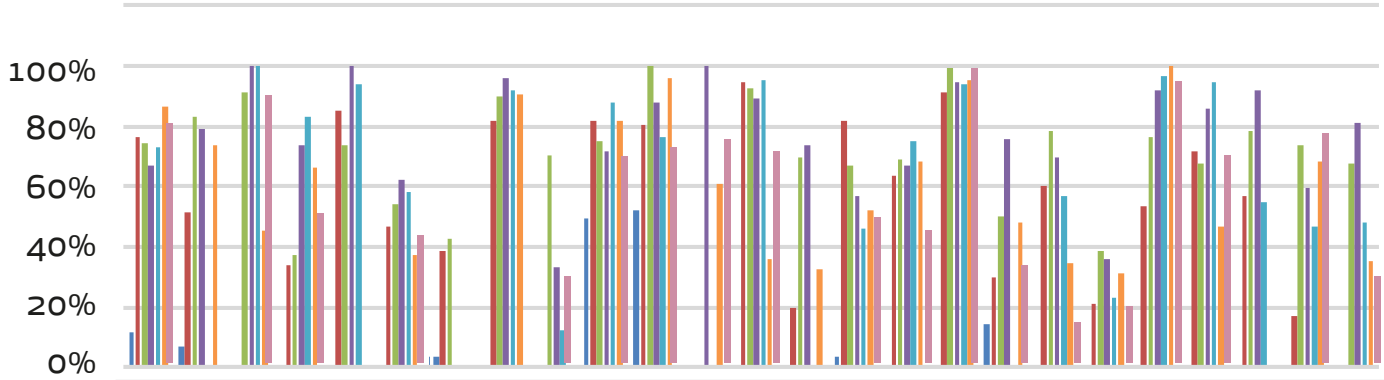
- Overall, IT companies removed **63.6%** of the content notified to them, while **36.4%** remained online. This result is slightly higher than the average of **62.5%** recorded in 2021, but lower than the peak score of **71%** in 2020.
- The results by IT company show that **only YouTube had a better score than last year**, while all the other platforms had a lower removal rate than in 2021.
- **Removal rates varied depending on the severity of hateful content.** On average, **69.6% of content calling for murder or violence against specific groups was removed,** while **content using defamatory words or pictures to name certain groups**, was removed in **59.1%** of the cases.
- The divergence in removal rates between content reported using trusted reporting channels as compared to channels available to all users was **25.4 percentage points, much higher than the 13.5** percentage points observed in 2021. This seems to suggest that there is a **growing difference of treatment** between the **notifications from general users and those sent through special channels for "trusted flaggers"**.
- IT companies were invited to make a self-assessment on the results of the exercise. Some of them reported cases in which they disagreed with the notifying organisations, i.e. where according to their assessment the content notified was not in violation of terms of services and/or local laws. For example, Facebook reported to disagree on **15.5%** of cases flagged to them and Instagram on **16.4%**. These percentages are slightly higher than in 2021, when Facebook had disagreement in **12%** and Instagram in **11.9%** of the cases. YouTube disagreed on **3%** of the cases. This shows the complexity of making assessments on hate speech content and calls for enhanced exchanges among trusted flaggers, civil society organisations and the content moderation teams in the IT companies.

YouTube removed **90.4%** of the content flagged, Facebook **69.1%**, TikTok **60.2%**, Instagram **58.4%** and Twitter **45.4%**. Except for YouTube, all the other platforms had a lower removal rate than in 2021, although often with minor variations (for example, Facebook removed **70.2%** of content in 2021 and Twitter **49.8%**).

Removals per IT Company



Rate of removals per EU country (in %)[3]



	AT	BE	BG	HR	CY	CZ	DK	EE	FI	FR	DE	EL	HU	IE	IT	LV	LT	NL	PL	PT	RO	SK	SI	ES	SE
1st monitoring (Dec 2016)	11.4	6.9					3.4			49.5	52				3.6			14.3					0%	0%	
2nd monitoring (May 2017)	76.1	51.2		33.6	84.8	46.5	38.9	81.8		82	80.1		94.5	20	81.7	63.3	90.9	29.5	60.3	21	53.3	71.4	56.8	17.2	
3rd monitoring (Jan 2018)	74	83	91.3	37.1	73.8	54.5	42.5	89.8	70.6	75	100		92.6	69.6	66.9	69	99	50	78.4	38.6	76.2	67.7	78.3	73.8	67.3
4th monitoring (Feb 2019)	66.7	78.9	100	73.4	100	62.5		95.8	39.1	71.8	87.6	100	89.2	73.5	56.6	66.9	94.9	75.8	69.7	35.7	92.2	85.8	91.7	59.7	81.3
5th monitoring (Jun 2020)	73.3		100	83.2	94.1	57.9		92	12.1	87.8	76.6		95		46.2	74.7	94.1		56.8	23.2	96.5	94.3	54.8	46.9	47.9
6th monitoring (Oct 2021)	86.8	73.5	45.2	66.3		37.4		90.6		81.5	95.9	60.6	36.1	32.3	52	68.4	95.4	47.9	34.5	31.2	100	46.5		68.2	35.4
7th monitoring (Nov 2022)	80.2		90.3	51.1		43.5			26	70.9	76.9	78.2	73.8		49	45.3	98.1	34.5	15.2	20.7	94.6	69.5		78.9	26.9

[3] The table does not reflect the prevalence on illegal hate speech online in a specific country and it is based on the number of notifications sent by each individual organisation. Estonia and Ireland are not included given the too low number of notifications (<20). No cases of hate speech were submitted from the following countries/ languages: Belgium, Slovenia, Malta, Cyprus, Luxembourg, and Denmark. Two organisations from the United Kingdom took part to the monitoring exercise: Tell Mama (80 cases), and Media Diversity Institute (56) with a total number of 136 cases submitted. Their work resulted on an average removal rate of 42,6% which is similar to the one recorded in 2021 (43%).

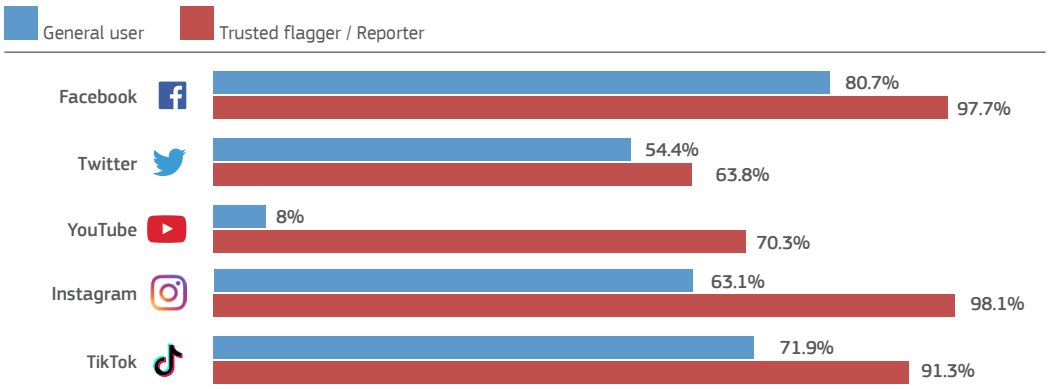
4. Feedback to users and transparency

- On average, the **IT companies provided feedback to 66.4%** of the notifications received. **This is higher than in the previous monitoring exercise (60.3%).**
- The Digital Service Act, which entered into force on 16 November 2022, highlights the need for of clearer 'notice-and-action' procedures, including transparency and feedback to users' notifications.

Facebook remains the platform that informs users most systematically (**84.9%** of notifications received feedback, similar to the 86.9% of 2021). TikTok and Instagram have improved their feedback to users, with **74.8%** and **72.6%** (28.7% and 41.9% in 2021, respectively). Twitter gave feedback to **57.1%** of the notifications (slightly higher than the 54.1% of 2021) and YouTube to **13.5%** (7.3% in 2021).

All platforms respond more frequently to notifications sent from the trusted flagger channels.

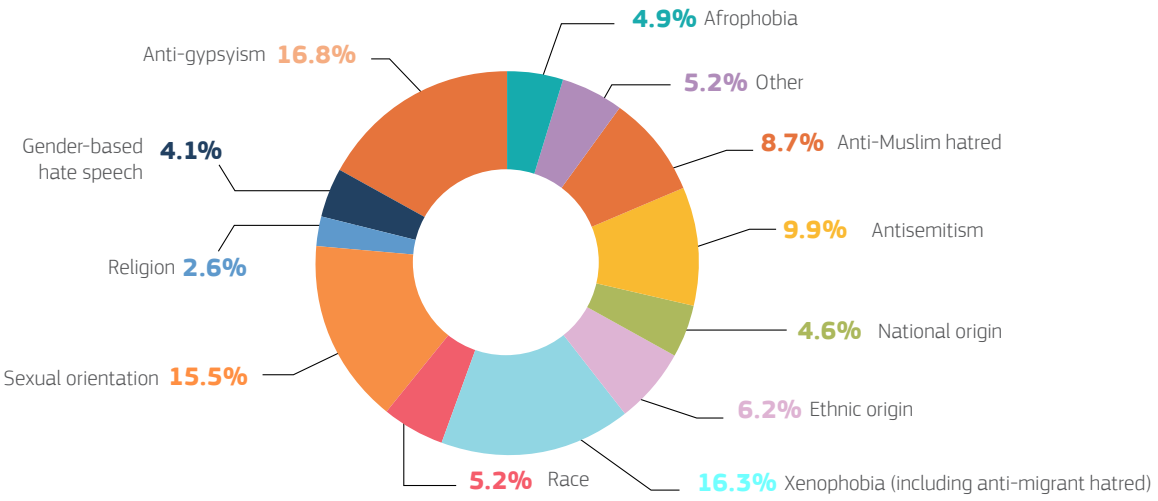
Feedback provided to different types of user



5. Grounds for reporting hatred

- In this monitoring exercise, **anti-gypsyism, xenophobia (including anti-migrant hatred) and sexual orientation** are the most commonly reported ground of hate speech.
- The data on grounds of hatred are only an indication and are influenced by the number of notifications sent by each organisation, as well as their field of work.

Grounds of hatred 2022



ANNEX

Methodology of the exercise

• The seventh exercise was carried out for a period of approximately 6 weeks (excluding public holidays), from 28 March to 13 May 2022, using the same methodology as the previous monitoring exercises.

• 33 civil society organisations and 3 public bodies (in France and Spain) reported on the outcomes of a total sample of 3634 notifications from 21 Member States.

• The figures do not intend to be statistically representative of the prevalence and types of illegal hate speech in absolute terms, and are based on the total number of notifications sent by the organisations.

• The organisations only notified the IT companies about content deemed to be "illegal hate speech" under national laws transposing the EU Council Framework Decision 2008/913/JHA on combating certain forms and expressions of racism and xenophobia by means of criminal law.

• Notifications were submitted either through reporting channels available to all users, or via dedicated channels only accessible to trusted flaggers/reporters.

• The organisations having the status of trusted flagger/reporter often used the dedicated channels to report cases which they previously notified anonymously (using the channels for all users) to check if the outcomes could diverge. Typically, this happened in cases when the IT companies did not send feedback to a first notification and content was kept online.

• The organisations participating in the seventh monitoring exercise are the following:

BELGIUM (BE)

CEJI - A Jewish contribution to an inclusive Europe

BULGARIA (BG)

Integro association

CZECH REPUBLIC (CZ)

In Iustitia
Romea

GERMANY (DE)

Jugendschutz.net
HateAid
FSM

ESTONIA (EE)

Estonian Human Rights Centre

FINLAND (FI)

University of Helsinki

IRELAND (IE)

INAR

GREECE (EL)

Greek Helsinki Monitor

SPAIN (ES)

Fundación Secretariado Gitano

Federación Estatal de Lesbianas, Gais, Transexuales y Bisexuales (FELGTB)

Spanish Observatory on Racism and Xenophobia (OBERAXE)

Spanish Ministry of Interior

FRANCE (FR)

Ligue Internationale Contre le Racisme et l'Antisémitisme (LICRA)

Platforme PHAROS

CROATIA (HR)

Centre for Peace Studies / Human Rights House Zagreb

ITALY (IT)

Ufficio Nazionale Antidiscriminazioni Razziali (UNAR)
CESIE
Centro Studi Regis
Amnesty International Italia
Associazione Carta di Roma

LATVIA (LV)

Mozaika

Latvian Centre for Human Rights

LITHUANIA (LT)

National LGBT Rights Oganisation (LGL)

HUNGARY (HU)

Háttér Society
Subjective Values Foundation

AUSTRIA (AT)

Zivilcourage und Anti-Rassismus-Arbeit (ZARA)

POLAND (PL)

Never Again Association

PORTUGAL (PT)

Associação ILGA Portugal

ROMANIA (RO)

Active Watch

SLOVAKIA (SK)

digiQ

SWEDEN (SE)

Institutet för Juridik och Internet

NETHERLANDS

INACH/Magenta
Meldpunt Internet Discriminatie (MiND