European Commission

Countering illegal hate speech online

5th evaluation of the Code of Conduct

Factsheet | June 2020

Didier Reynders
Commissioner for Justice

Directorate-General for Justice and Consumers

The fifth evaluation on the **Code of Conduct on Countering Illegal Hate Speech Online** shows that the Code continues to deliver positive results. On average **90% of the notifications are reviewed within 24 hours** and **71% of the content is removed.** While the average removal rate is stable with respect to previous monitoring exercises, some divergences exist among the platforms. Most of the IT companies **must improve their feedback to users' notifications.**

1. Notifications of illegal hate speech

- **39 organisations** from 23 Member States and the United Kingdom **sent notifications relating to hate speech deemed illegal to the IT companies** during a period of 6 weeks (4 November to 13 December 2019). In order to establish trends, this exercise used the same methodology as the previous monitoring rounds (see Annex).
- A total of **4364 notifications** were submitted to the IT companies taking part in the Code of Conduct.
- **2 513 notifications** were submitted through the reporting **channels available to general users**, while **1851** were submitted through **specific channels available only to trusted flaggers/reporters**.
- **Facebook received the largest amount of notifications (2 348)**, followed by Twitter **(1396)**, YouTube **(464)** and Instagram **(109)**. Jeuxvideo.com **(40)** and Dailymotion **(7)**[1], were tested too.
- In addition to flagging the content to IT companies, the organisations taking part in the monitoring exercise submitted **475 cases of hate speech** to the police, public prosecutor's bodies or other national authorities.

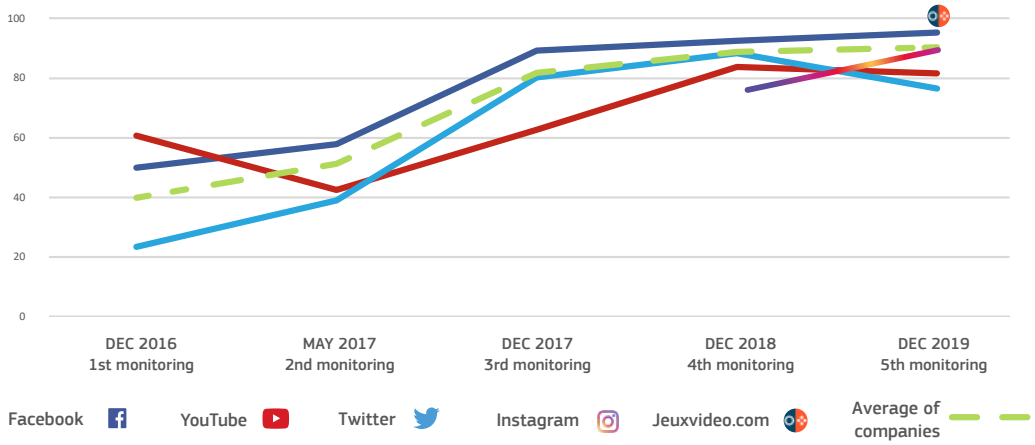
[1] Given the very limited amount of cases, the results of Dailymotion will not be quantified in % points.

2. Time of assessment of notifications

- In **90.4% of the cases** the IT companies assessed the notifications **in less than 24 hours,** an additional **4.9%** in less than 48 hours, **4.3%** in less than a week and in **0.4%** of cases it took more than a week.
- **The target of reviewing the notifications within one day is fully met** by all the IT companies and the **trend of progress** compared to the previous monitoring exercise **continues** (it was **89%** in 2019).

> Facebook assessed notifications in less than 24 hours in **95.7%** of the cases and 3.**4%** in less than 48 hours. The corresponding figures for YouTube are **81.5%** and **8.7%** and for Twitter **76.6%** and **8.7%**, respectively. Instagram's performance is very positive, with **91.8%** of notifications assessed in less than 24 hours, and Jeuxvideo.com did so in all the cases.

Percentage of notifications assessed within 24 hours - Trend over time

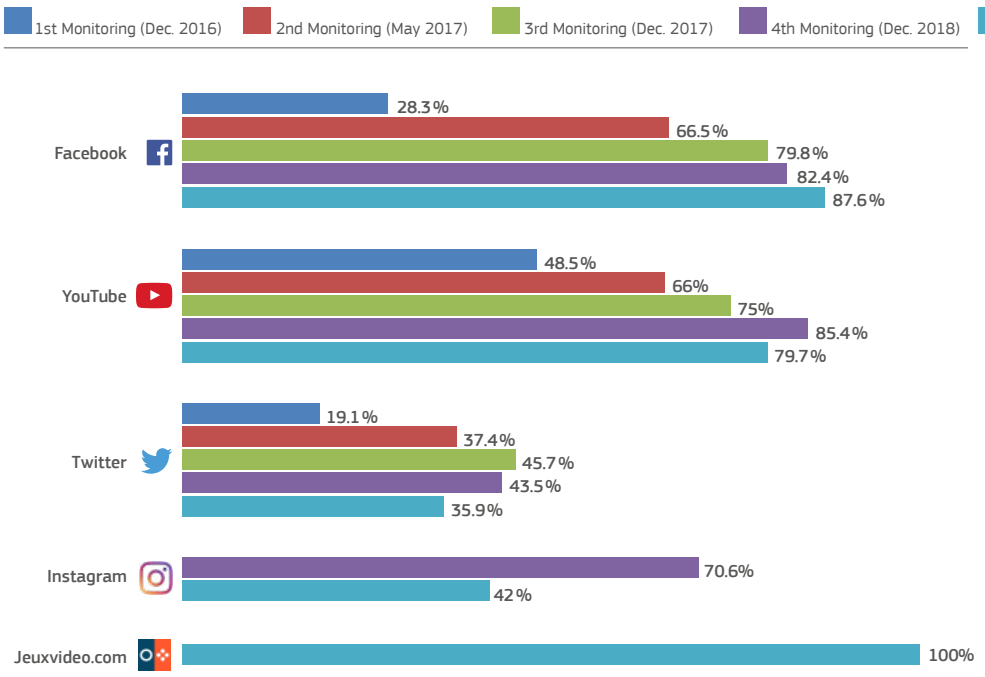


3. Removal rates

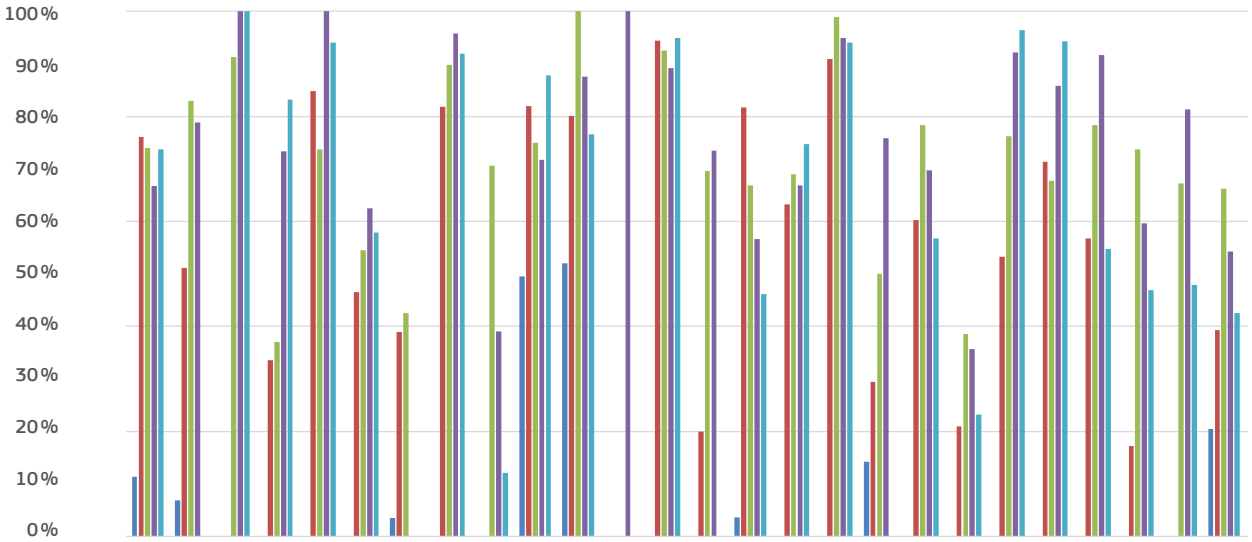
- Overall, **IT companies removed 71% of the content** notified to them, while **29%** remained online. This is **in line** with the average of **71.7%** recorded one year ago.
- **Removal rates varied depending on the severity of hateful content.** On average, **83.5% of content calling for murder or violence of specific groups was removed,** while content using defamatory words or pictures to name certain groups was removed in **57.8%** of the cases. This suggest that the reviewers assess the content **scrupulously** and with **full regard** to protected speech.
- The divergence in removal rates of content reported using trusted reported channels as compared to channels available to all users **was 16.2** percentage points. **This difference has increased more than three times in percentage points compared to 2018 (4.8%).** This seems to suggest that **notifications from general users are often treated differently** than those sent through special channels for "trusted flaggers"
- IT companies were invited to make a self-assessment on the results of the exercise. They reported cases in which they disagreed with the notifying organisations, i.e. where in their assessment the content notified was not in violation of terms of services and/or local laws.

> Facebook removed **87.6%** of the content, YouTube **79.7%**, and Twitter **35.9%**. Facebook made further progress on removals compared to last year. YouTube remains at high standards while Twitter is not in target and the removal rate is lower than in 2019. Jeuxvideo.com removed all flagged content and Instagram **42%**.

Removals per IT Company



Rate of removals per EU country (in %)[2]



	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	UK
1st monitoring (Dec. 2016)	11.4	6.9					3.4			49.5	52				3.6			14.3					0%	0%		20.5
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		39.3
3rd monitoring (Dec. 2017)	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	66.3
4th monitoring (Dec. 2018)	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	54.3
5th monitoring (Dec. 2019)	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	42.5

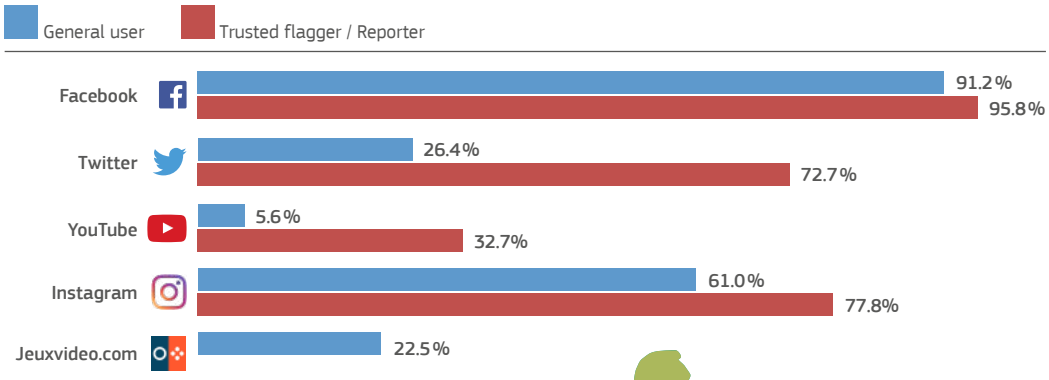
[2] The table does not reflect the prevalence on illegal hate speech online in a specific country and it is based on the number of notifications sent by each individual organisation., Belgium, Greece and Ireland are not included given the too low number of notifications made to companies (<20). For Malta, Luxembourg, the Netherlands and Denmark the organisations did not submit cases for this exercise.

4. Feedback to users and transparency

- On average, the **IT companies responded with a feedback to 67.1%** of the notifications received. **This is slightly higher than in the previous monitoring exercise (65.4%).**
- **A correlation can be observed between systematic feedback to users, the swift review of notifications and the effective removal when needed.** The 2018 European Commission Recommendation on measures to effectively tackle illegal content online highlights the importance of clearer 'notice-and-action' procedures including transparency and feedback to users' notifications.

> Only Facebook is informing users systematically (**93.7%** of notifications received feedback). Instagram gave feedback to **62.4%** of the notifications, Twitter to **43.8%** and YouTube only to **8.8%**. Jeuxvideo.com sent feedback to **22.5%** of the notifications.
>
> While Facebook is the only company informing consistently both trusted flaggers and general users, Twitter, YouTube and Instagram provide feedback more frequently when notifications come from trusted flaggers.

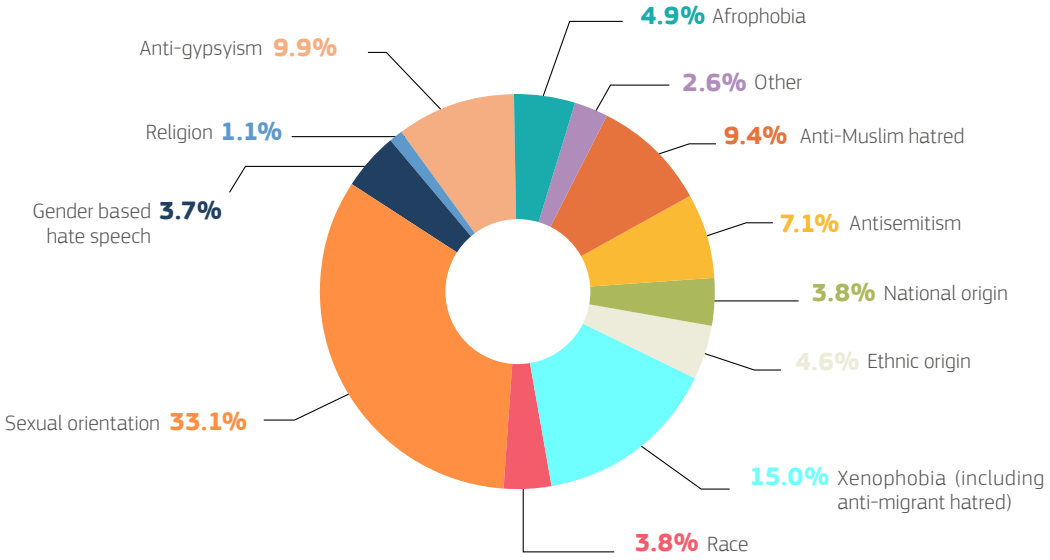
Feedback provided to different types of user



5. Grounds for reporting hatred

- In this monitoring exercise, **sexual orientation is the most commonly reported ground of hate speech (33.1%)** followed by xenophobia (including anti-migrant hatred) **(15%)** and anti-gypsyism **(9.9%)**,
- 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[3].

Grounds of hatred 2019



[3] In this monitoring round, organisations working on LGBTI rights have been more active in flagging content, in relative terms.

ANNEX

Methodology of the exercise

- The fifth exercise was carried out for a period of 6 weeks, from 4 November to 13 December 2019, using the same methodology as the previous monitoring exercises.
- 34 organisations and 5 public bodies (in Belgium, France, Spain, and Finland) reported on the outcomes of a total sample of 4364 notifications from all the Member States (and plus the United Kingdom), except for Luxembourg, the Netherlands, Malta and Denmark.
- 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 fifth monitoring exercise are the following:

COUNTRY	N° OF CASES
BELGIUM (BE)	
CEJI - A Jewish contribution to an inclusive Europe	12
Centre interfédéral pour l'égalité des chances (UNIA)	14
BULGARIA (BG)	
Integro association	101
CZECH REPUBLIC (CZ)	
In Iustitia	100
Romea	34
GERMANY (DE)	
Freiwillige Selbstkontrolle Multimedia-Diensteanbieter e.V. (FSM e.V.)	38
Jugendschutz.net	26
ESTONIA (EE)	
Estonian Human Rights Centre	100
IRELAND (IE)	
ENAR Ireland	15
GREECE (EL)	
SafeLine / Forth	4
SPAIN (ES)	
Fundación Secretariado Gitano	108
Federación Estatal de Lesbianas, Gais, Transexuales y Bisexuales (FELGTB)	99
Spanish Observatory on Racism and Xenophobia (OBERAXE)	102
Spanish Ministry of Interior	45
FRANCE (FR)	
Ligue Internationale Contre le Racisme et l'Antisémitisme (LICRA)	98
Platforme PHAROS	81
CROATIA (HR)	
Centre for Peace Studies	101
ITALY (IT)	
Ufficio Nazionale Antidiscriminazioni Razziali (UNAR)	212
CESIE	100
Centro Studi Regis	34
Amnesty International Italia	112
Associazione Carta di Roma	48

COUNTRY	N° OF CASES
CYPRUS (CY)	
Aequitas	101
LATVIA (LV)	
Mozaika	101
Latvian Centre for Human Rights	77
LITHUANIA (LT)	
National LGBT Rights Oganisation (LGL)	1002
HUNGARY (HU)	
Háttér Society	99
AUSTRIA (AT)	
Zivilcourage und Anti-Rassismus-Arbeit (ZARA)	99
POLAND (PL)	
HejtStop / Projekt: Polska	112
PORTUGAL (PT)	
Associação ILGA Portugal	94
ROMANIA (RO)	
Active Watch	86
SLOVENIA (SI)	
sCAN Spletno oko	83
SLOVAKIA (SK)	
digiQ	141
FINLAND (FI)	
Finnish Police Academy	33
SWEDEN (SE)	
Institutet för Juridik och Internet	73
UNITED KINGDOM (UK)	
Media Diversity Insitute	80
Galop	100
Community Security Trust	100
Tell Mama/Faith Matters	4