European
Commission

4th evaluation confirms self-regulation works

Code of Conduct on countering illegal hate speech online

Factsheet | **January 2019**

Věra Jourová
Commissioner for Justice, Consumers and Gender Equality

Directorate-General for Justice and Consumers

The fourth evaluation on the *Code of Conduct on Countering Illegal Hate Speech Online* confirms continuous progress on the swift removal of illegal hate speech. While the fight against hate speech and toxic narratives online needs to be continued and further strengthened, the Code is delivering on its key commitments. It proves to be an effective tool to face the challenge.

Today, all IT Companies fully meet the target of reviewing the majority of the notifications within **24 hours**, reaching an average of more than 89%. These results also include Instagram and Google+ which joined in 2018. This is a significant increase from when the Code was launched back in 2016 (40% within 24 hours).

On average, IT companies **are removing 72 % of the illegal hate speech notified** to them. This is estimated to be satisfactory removal rates, as some of the content flagged by users could relate to content that is not illegal. In order to protect freedom of speech only content deemed illegal should be removed.

Key figures

1. Notifications of illegal hate speech

> 39 organisations from 26 Member States (all except Luxembourg and Denmark) sent notifications relating to hate speech deemed illegal to the IT companies during a period of 6 weeks (5 November to 14 December 2018). In order to establish trends, this exercise used the same methodology as the previous monitoring rounds (see Annex).

> A total of 4 392 notifications were submitted to the IT companies taking part in the Code of Conduct. This represents a steady increase compared to the previous exercises.

> 2 748 notifications were submitted through the reporting channels available to general users, while 1644 were submitted through specific channels available only to trusted flaggers/reporters.

> Facebook received the largest amount of notifications (1 882), followed by Twitter (1314) and YouTube (889). This breakdown is similar to previous exercises. Instagram (279) and Google+ (28), which have joined the Code of conduct in early 2018, were tested too. Microsoft did not receive any notification.

> In addition to flagging the content to IT companies, the organisations taking part in the monitoring exercise submitted 503 cases of hate speech to the police, public prosecutor's bodies or other national authorities.

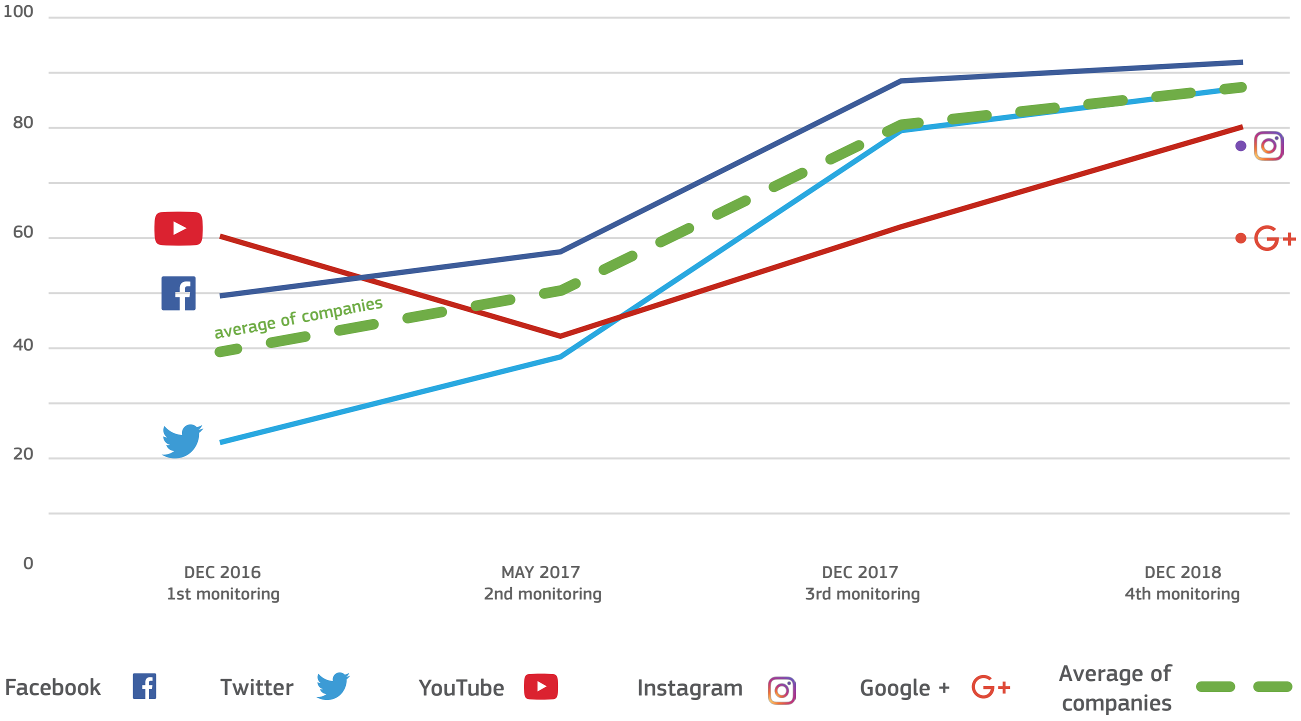
2. Time of assessment of notifications

> In **88.9 % of the cases** the IT companies assessed the notifications **in less than 24 hours**, an additional 6.5 % in less than 48 hours, 3.9 % in less than a week and in 0.7 % of cases it took more than a week.

> Facebook assessed the notifications in less than 24 hours in 92.6 % of the cases and 5.1 % in less than 48 hours. The corresponding figures for YouTube are 83.8 % and 7.9 % and for Twitter 88.3 % and 7.3 %, respectively. Instagram's performance is positive, 77.4 % of notifications were assessed in less than 24 hours, while Google+ did so in 60% of the cases[1].

> The target of reviewing the notifications within one day is fully met by all the IT companies and there has been additional progress compared to the previous monitoring exercise (81.7%).

Rate of notifications reviewed within 24 hours since the launch of the Code of Conduct



[1] The figures for Google+ are based on a significantly lower number of cases compared to the other IT companies.

3. Removal rates

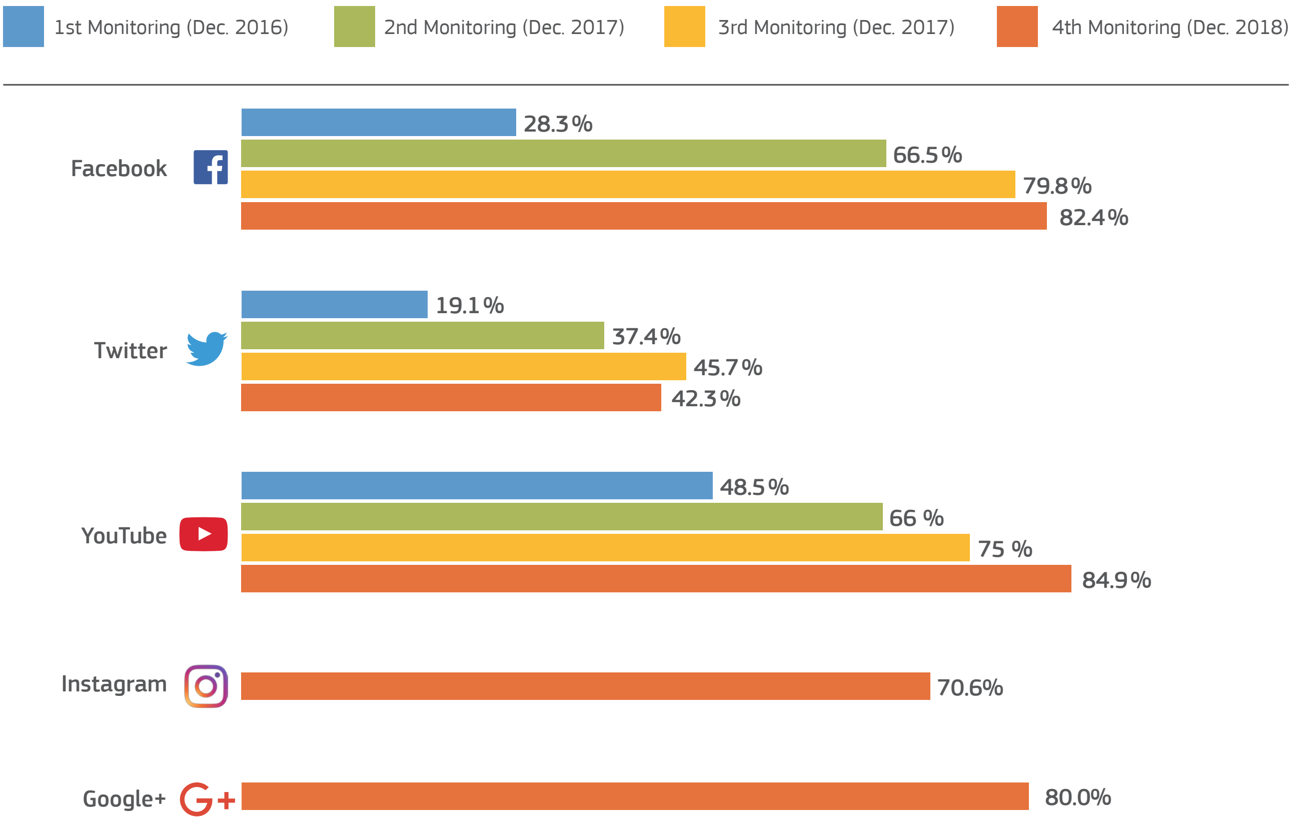
> Overall, **IT companies removed 71.7 % of the content** notified to them, while 28.3 % remained online. This represents a small increase compared to the 70% one year ago.

> YouTube removed 85.4 % of the content[2], Facebook 82.4 % and Twitter 43.5 %. Both Facebook and, especially, YouTube made further progress on removals when compared to last year. Twitter, while remaining in the same range as in the last monitoring, has slightly decreased its performance. Google+ removed 80% of the content and Instagram 70.6%.

> Removal rates varied depending on the severity of hateful content. On average, 85.5 % 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 58.5% of the cases. This suggest that the reviewers assess the content scrupulously and with full regard to protected speech.

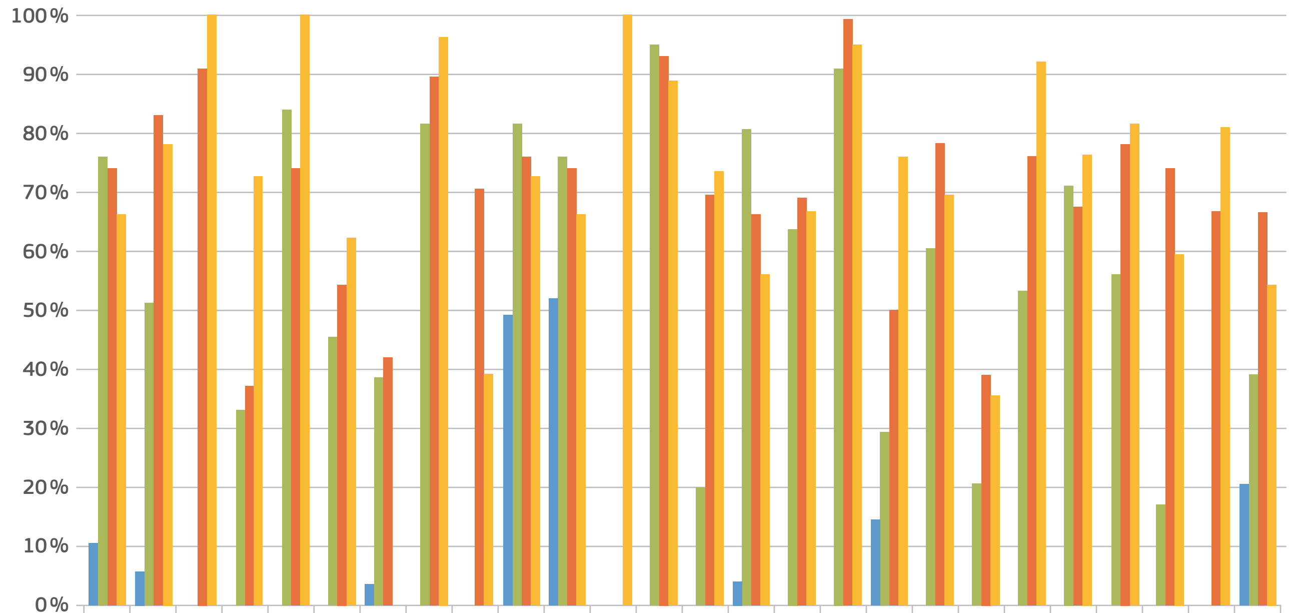
> The divergence in removal rates of content reported using trusted reported channels as compared to channels available to all user was only 4.8%. This difference was more than twice as high in December 2017 (10.5%).

Removals per IT Company



[3] YouTube has also limited the features of an additional 23 videos: this implies that while not being removed, a video may not be liked, commented, or shared and does not appear in searches.

Rate of removals per EU country[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	UK
🟦 1st monitoring (Dec. 2016)	11.4	6.9					3.4			49.5	52.0				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		82.0	80.1		94.5	20.0	81.7	63.3	90.9	29.5	60.3	21.0	53.3	71.4	56.8	17.2		39.3
🟥 3rd monitoring (Dec. 2017)	74.0	83.0	91.3	37.1	73.8	54.5	42.5	89.8	70.6	75.0	100.0		92.6	69.6	66.9	69.0	99.0	50.0	78.4	38.6	76.2	67.7	78.3	73.8	67.3	66.3
🟨 4th monitoring (Dec. 2018)	66.7	78.9	100.0	73.4	100.0	62.5		95.8	39.1	71.8	87.6	100.0	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

4. Feedback to users and transparency

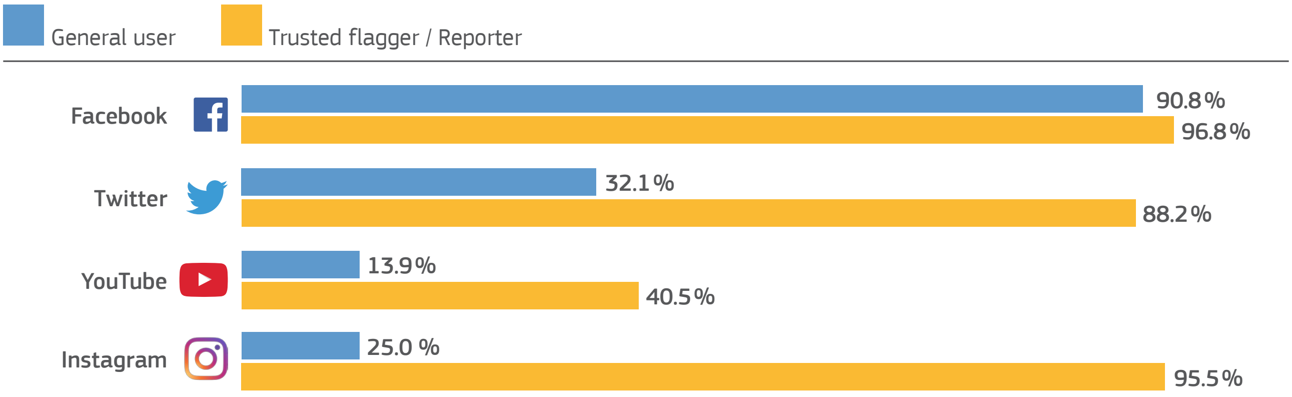
> On average, the IT companies responded with a feedback to 65.4 % of the notifications received. This is slightly lower than in the previous monitoring exercise (68.9%). Only Facebook is informing users systematically (92.6% of notifications received a feedback), Twitter gave feedback to 60.4% of the notifications and YouTube only to 24.6%. The corresponding figures in December 2017 were 94.8%, 70.4%, and 20.8% respectively.

> While Facebook is the only company informing consistently both trusted flaggers and general users, Twitter and YouTube provide feedback more frequently when notifications come from trusted flaggers (88.2% and 40.5% respectively).

> Instagram sent feedback to 95.5% of the notifications from trusted flaggers and to 25% of those from general users. Google+ did not send feedback to any notification.

[3] The table does not reflect the global issue on illegal hate speech online in a specific country and it is based on the number of notifications sent by each individual organisation. Malta and Greece are not included given the too low number of notifications made to companies (<20). For Luxembourg, no organization participated to this exercise.

Feedback provided to different types of user

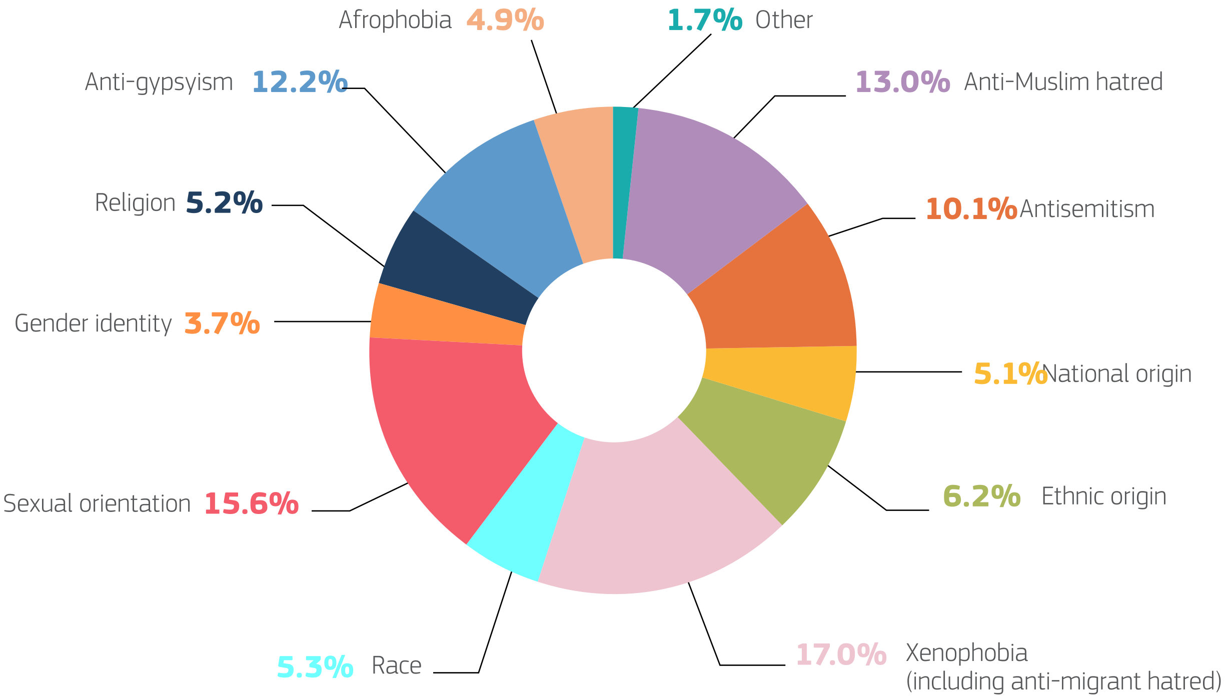


5. Grounds for reporting hatred

> Xenophobia (including anti-migrant hatred) is the most commonly reported grounds of hate speech (17%) followed by sexual orientation (15.6%) and anti-Muslim hatred (13 %).

> The results, which are in line with the trends in December 2017, confirm the predominance of racist hatred against ethnic minorities, migrants and refugees. Data on grounds of hatred are only an indication of trends and may be influenced by the field of activity of the organisations participating to the monitoring exercise.

Grounds of hatred



ANNEX

Methodology of the exercise

- The fourth exercise was carried out for a period of 6 weeks, from 5 November to 14 December 2018, using the same methodology as the previous monitoring exercises.
- 35 organisations and 4 public bodies (in France, Spain, UK and Finland) reported on the outcomes of a total sample of notifications from all the Member States except for Luxembourg and Denmark. An additional 26 cases were reported to other platforms.
- 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 fourth monitoring exercise are the following:

COUNTRY	N° OF CASES
BELGIUM (BE)	
CEJI - A Jewish contribution to an inclusive Europe	14
Centre interfédéral pour l'égalité des chances (UNIA)	38
BULGARIA (BG)	
Integro association	101
CZECH REPUBLIC (CZ)	
In Iustitia	101
Romea	35
GERMANY (DE)	
Freiwillige Selbstkontrolle Multimedia-Diensteanbieter e.V. (FSM e.V.)	89
Jugendschutz.net	104
ESTONIA (EE)	
Estonian Human Rights Centre	96
IRELAND (IE)	
ENAR Ireland	67
GREECE (EL)	
SafeLine / Forth	30
SPAIN (ES)	
Fundación Secretariado Gitano	109
Federación Estatal de Lesbianas, Gais, Transexuales y Bisexuales (FELGTB)	98
Spanish Observatory on Racism and Xenophobia (OBERAXE)	284
FRANCE (FR)	
Ligue Internationale Contre le Racisme et l'Antisémitisme (LICRA)	111
Ligue Internationale Contre le Racisme et l'Antisémitisme (LICRA)	111
CROATIA (HR)	
Centre for Peace Studies	91
ITALY (IT)	
Ufficio Nazionale Antidiscriminazioni Razziali (UNAR)	434
CESIE	111
Centro Studi Regis	87
CYPRUS (CY)	
Aequitas	101

COUNTRY	N° OF CASES
LATVIA (LV)	
Mozaika	58
Latvian Centre for Human Rights	85
LITHUANIA (LT)	
National LGBT Rights Oganisation (LGL)	316
HUNGARY (HU)	
Háttér Society	71
MALTA (MT)	
Malta LGBTIQ Right Movement (MGRM)	5
NETHERLANDS (NL)	
Meldpunt Internet Discriminatie (MiND)	1
INACH / Magenta Foundation	100
AUSTRIA (AT)	
Zivilcourage und Anti-Rassismus-Arbeit (ZARA)	102
POLAND (PL)	
HejtStop / Projekt: Polska	143
PORTUGAL (PT)	
Associação ILGA Portugal	98
ROMANIA (RO)	
Active Watch	153
SLOVENIA (SI)	
Spletno oko	100
SLOVAKIA (SK)	
digiQ	106
FINLAND (FI)	
Finnish Police Academy	69
SWEDEN (SE)	
Institutet för Juridik och Internet	64
UNITED KINGDOM (UK)	
True Vision	1
Galop	100
Community Security Trust	136
Tell Mama/Faith Matters	3