Code of Conduct on countering illegal hate speech online: First results on implementation

Factsheet | **December 2016**

Věra Jourová
Commissioner for Justice, Consumers and Gender Equality

Directorate-General for Justice and Consumers

1. Introduction

On the 31 May 2016, the Commission presented with Facebook, Microsoft[1], Twitter and YouTube ("the IT Companies") a "Code of conduct on countering illegal hate speech online". The main commitments are:

a) The IT Companies to have in place clear and effective processes to review notifications regarding illegal hate speech on their services so they can remove or disable access to such content. The IT Companies to have in place Rules or Community Guidelines clarifying that they prohibit the promotion of incitement to violence and hateful conduct.

b) Upon receipt of a valid removal notification, the IT Companies to review such requests against their rules and community guidelines and, where necessary, national laws transposing the Framework Decision 2008/913/JHA, with dedicated teams reviewing requests.

c) The IT Companies to review the majority of valid notifications for removal of illegal hate speech in less than 24 hours and remove or disable access to such content, if necessary.

The IT Companies and the European Commission agreed to assess the public commitments in the code of conduct on a regular basis, including their impact.

To ensure an effective measuring of progress, the Commission's sub-group on countering hate speech online agreed, on 5 October 2016, on a common methodology to assess the reactions of IT Companies upon notification of illegal hate speech. It was also agreed that the preliminary results of this monitoring exercise would be reported to Member States, IT Companies and civil society organisations in the framework of the High Level Group on combatting Racism, Xenophobia and other forms of intolerance.

For 6 weeks, 12 organisations based in 9 different Member States applied the common methodology. The organisations notified alleged illegal hate speech online (as defined in national criminal codes transposing the Framework Decision) to the IT Companies and used a commonly agreed template to record, when possible, the rates and timings of take-downs in response to the notifications.

[1] Microsoft-hosted consumer services, as relevant

The monitoring exercise is a continuous process. These initial data constitute a baseline and a first valuable indication of the current situation. A second monitoring cycle will be carried out during 2017 to observe trends.

2. Methodology of the exercise

12 organisations located in 9 Member States volunteered to test the reactions of IT Companies upon notification of alleged illegal hate speech content and to record their response. Participating organisations are listed in the table below.

- The exercise was carried out during a period of six weeks (from 10 October to 18 November 2016)[2].
- The 12 organisations reported a sample of 600 notifications in the following Member States: Austria, Belgium, Denmark, France, Germany, Italy, The Netherlands, Spain, United Kingdom.
- The organisations notified content to IT Companies, by using dedicated reporting channels ("Trusted reporters/flaggers") or through the tools available to normal users. Trusted flaggers, trusted reporters or equivalent mechanism, refers to the status given to certain organisations which allows them to report illegal content through a special reporting system or channel, which is not available to normal users.
- The organisations taking part in the monitoring exercise are the following:

participating organisations
Freiwillige Selbstkontrolle Multimedia-Diensteanbieter e.V. (FSM e.V. – Germany) – 122 cases **jugendschutz.net** (INACH member - Germany) - 55 cases
Ufficio Nazionale Antidiscriminazioni Razziali (UNAR – Italy) – 110 cases
Zivilcourage und Anti-Rassismus-Arbeit (ZARA, INACH member - Austria) – 88 cases
Community Security Trust (CST, INACH member - United Kingdom) – 78 cases
International League Against Racism And Antisemitism (LICRA, INACH member – France) -74 cases
Center for Forebyggelse af Eksklusion - Anmeldhad.dk (CFE – Denmark) – 29 cases
A Jewish Contribution to an inclusive Europe (CEJI – EU umbrella - Belgium) – 16 cases **Centre Interfédéral pour l'égalité des chances** (UNIA, INACH member – Belgium) – 13 cases
Movimiento contra la intolerancia (MCI, INACH member – Spain) – 8 cases
INACH-Magenta Foundation (EU umbrella organisation - the Netherlands) - 5 cases **Meldpunt Internet Discriminatie** (MiND, INACH member -The Netherlands) – 2 cases

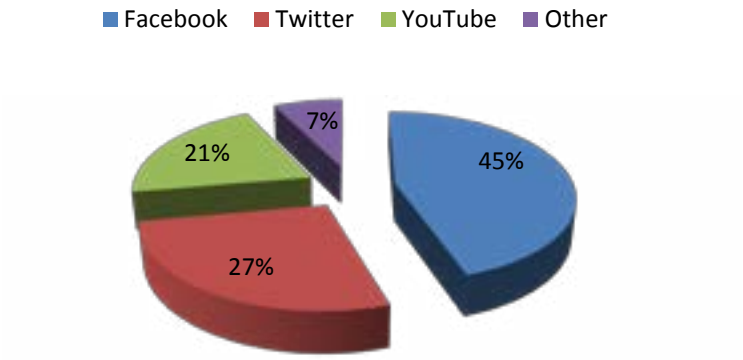
- Differences in the number of notifications made do not reflect the global issue of illegal hate speech online in a specific country. Rather the differences correspond to the resources invested by the organisations involved and whether social platforms were actively scanned for illegal hate speech online or only acting upon citizens' complaints.

[2] Any data or information provided outside the monitoring period has not been taken into account for the results of the exercise.

3. Illegal hate speech notifications to IT Companies

- From the total notifications of illegal hate speech content: 270 have been made to Facebook, 163 to Twitter and 123 to YouTube. No notification has been made to Microsoft.

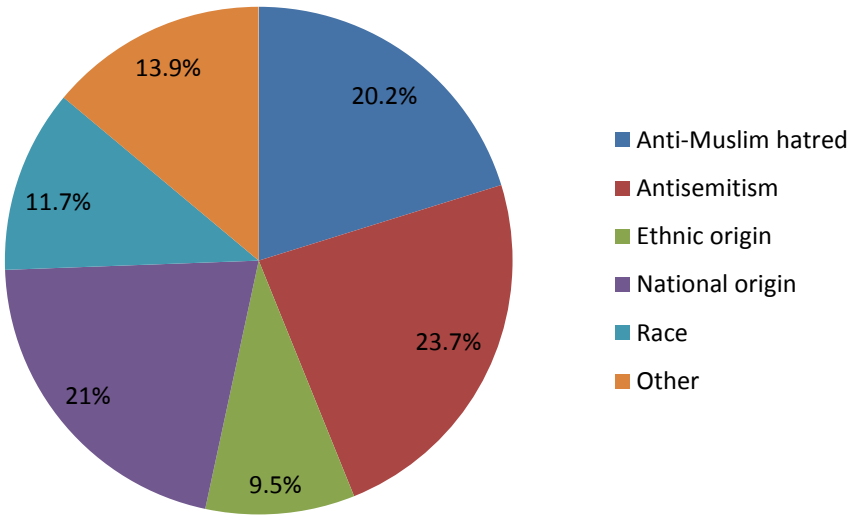
Number of notifications per IT company



These figures correspond to the notifications made by the organisations, which may not reflect the overall issue and amount of illegal hate speech on each of the IT Company's platforms. The category "other" refers to notifications sent to other IT Companies or social platforms, which are not covered by the Code of conduct on countering illegal hate speech online.

- The grounds for reporting hatred were the following: race, colour, national origin, ethnic origin, descent, religion, anti-Muslim hatred, Antisemitism, sexual orientation or gender-related hatred. A large number of cases corresponded to some form of anti-migrant speech identified on the grounds of anti-Muslim hatred, ethnic origin or race, depending on the context of the message.

Notified content by ground

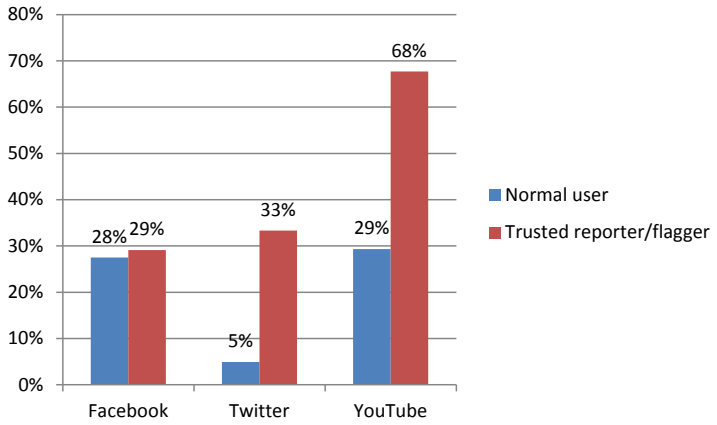


Data on grounds of hatred is an indication of trends and may be influenced by the field of activity of the organisations. For example, three organisations participating in the exercise are specialised in monitoring illegal Antisemitic hate speech online.

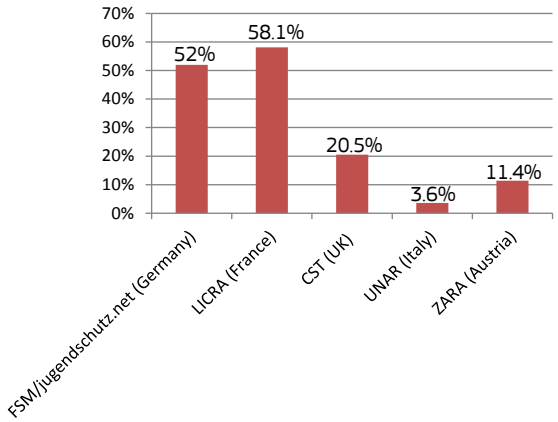
4. Notifications and removals

- Out of 600 notifications, 270 were made as "Trusted flaggers", 330 as "normal users".
- Overall, in 169 cases (28.2%) the content was removed.
- Facebook removed the content in 28.3% of cases, Twitter in 19.1% and YouTube in 48.5%.
- The reactions by Twitter and YouTube upon notification of illegal hate speech seem to diverge depending on the source used to notify content (trusted reporter/flagger system vs normal user tools). The ratios of removal for Facebook are similar, whether the user notifies the content through the trusted reporter channel or the normal tool.

% of removals per source of notification



% of removals on total notifications



The graph only includes the countries and organisations who reported more than 50 notifications to the IT companies.

5. Time spent by the IT Companies to deal with notifications of illegal hate speech

- Data recorded show that in 40% of the cases IT Companies reviewed the notification on the same day (less than 24h) and in 43% of these cases on the day after (less than 48h).
- Facebook assessed the notified content in less than 24 hours in 50% of the cases and in 41.9% of the cases in less than 48 hours. The corresponding figures for YouTube are 60.8% and 9.8% and for Twitter 23.5% and 56.1%, respectively.

Code of conduct on countering illegal hate speech online
http://ec.europa.eu/justice/fundamental-rights/files/hate_speech_code_of_conduct_en.pdf