HOW TO IMPROVE REPORTING INSTANCES RELATED TO CYBER-HATE PHENOMENA

By Rianne Pattipeilohij and Camille Lhopitault

Project Research - Report - Remove: Countering Cyber Hate Phenomena

INACH

Supported by the Rights, Equality and Citizenship (REC) Programme of the European Union

Executive Foreword

This publication was written within the framework of the ***Research – Report – Remove: Countering Cyber Hate Phenomena*** project of the International Network Against Cyber Hate (INACH); funded by the European Commission Directorate-General for Justice and Consumers. The duration of the project is 2016-2017, and its aim is to study, document and report on online hate speech in a comparative and comprehensive way; and to establish structures for a transnational complaints system for instances of cyber hate.

Hate speech is intentional or unintentional public discriminatory and/or defamatory statements; intentional incitement to hatred and/or violence and/or segregation based on a person's or a group's real or perceived race, ethnicity, language, nationality, skin colour, religious beliefs or lack thereof, gender, gender identity, sex, sexual orientation, political beliefs, social status, property, birth, age, mental health, disability, disease.

This report was completed with the participation of the different members of the Network and partners in the project, namely the Zivilcourage und Anti-Rassismus-Arbeit (ZARA) from **Austria**, the Movimiento contra la Intolerancia (MCI) from **Spain**, jugendschutz.net from **Germany**, the Ligue Internationale Contre le Racisme et l'Antisémitisme (LICRA) from **France**, the Inter-Federal Centre For Equal Opportunities and Opposition to Racism from **Belgium** (now called Unia), and the Magenta Foundation from the **Netherlands** (MDI); who provided most of the data this report is based upon.

Legal Disclaimer

This publication has been produced with the financial support of the Rights, Equality and Citizenship (REC) Programme of the European Union. The contents of this publication are the sole responsibility of the International Network Against Cyber Hate and can in no way be taken to reflect the views of the European Commission.

CONTENT

INTRODUCTION

Internet and social media have opened up new arenas without borders for exchanging opinions and for developing access to freedom of speech. However, at the same time, online hate speech has been spread widely and frequently on the platforms. Conspiracy theories, violent, hateful and racist speech may cause fear, violence, and social conflict. As a consequence it can be the reason why people withdraw from the public debate.

Civil society, especially NGOs have been sounding the alarm for a long time. Their involvement in countering cyber hate phenomena is becoming increasingly important. The use of the existing tools of reporting proposed by the IT companies and the development of new strategies for tackling hate speech have been the key priorities. By monitoring and reporting, civil society and NGOs are playing a significant role for weakening the online presence of violent and hateful speech.

For improving this effort, new steps have been taken in Europe, as for example the Code of Conduct between major IT companies and European Commission[1]. This Code has been welcomed, but the efforts must be maintained. To help those efforts, INACH would like to give some suggestions based on two concepts: transparency and simplification.

Transparency: "*the transparency of a process, situation, or statement is its quality of being easily understood or recognized*" (Collins dictionary).

Simplification: *"The process of making something simpler or easier to do or understand"* (Oxford dictionary).

Based on the EC monitoring sessions, in this recommendation report we decided to focus on the three major social networks and platforms: Facebook, Twitter and YouTube.

[1] European Commission and IT Companies announce Code of Conduct on illegal online hate speech, Brussels, 31 May 2016: http://europa.eu/rapid/press-release_IP-16-1937_en.htm

TRANSPARENCY OF THE SOCIAL NETWORKS REPORTING SYSTEMS

Provide sufficient information on how takedown procedures are executed and on what basis[2].

The power that social media has regarding the billions of users increases, which makes it more and more important for them to communicate on the mechanisms of their reporting systems. It is quite important for civil society, NGOs as well as citizens, media and institutional authorities to have global access to the information related to the treatment of hate speech on social media platforms.

Transparency on global data

Even if NGOs, for example INACH, do decide to aggregate their data regarding reported content and removed content, at this moment it is still impossible to have an objective observation of the amount of hate speech content per country or on a European level. And although this information is key in tackling cyber hate, only the social media companies have access to this information.

We would like to ask the companies to produce annual reports with their global data including:

- Number of reported content per country and on a European and global level;
- Global report of removed content per category of reporting;
- Removal rate per category of reporting.

Transparency about the actors of the reporting system

There is a lack of information regarding the number of people working with the reporting systems, the training and credentials of these people, if they are part of the companies or depending on subcontracting companies. To work on a safer online environment together, it would be recommended that trusted flaggers have a clear knowledge about the actors working within the reporting system at each company:

- Who is in charge of this reporting system?
- How many people are assigned to the reporting system?

[2] INACH Annual Conference 2017, Vienna, 12 October 2017, recommendations presented to the OSCE INTERNET FREEDOM CONFERENCE - THE ROLE AND RESPONSIBILITIES OF INTERNET INTERMEDIARIES held at the Hofburg in Vienna: http://www.inach.net/fileadmin/user_upload/INACH2017Final_Recommendations-OSCE.pdf

- How many reports do employees have to handle each day? How much time does it take to evaluate reported content?
- What kind of legal criteria are used for removing content? Are these criteria internal, international, European or national?
- What kind of training is given to employees to work with the reporting system?
- How slang, local dialects and/or other informal language is dealt with in the reporting system?

Transparency about the procedures of removal

As a user of the reporting systems, INACH would like to underline the inconsistency in the removal or blocking of reported content. The measures to stop online hate speech from spreading are plentiful: e.g. removing the content, block the content by using country specific 'country-blocking', block content for a certain period of time, hide content behind a warning button saying 'inappropriate or offensive content', limiting the features available to certain content or suspend user accounts. Although it is good to see that social media works hard on improving the online environment, there are no clear guidelines in the community standards explaining hot it is decided when to use which of these different blocking techniques or removals. To improve the collaboration and relationship between flaggers and social media it is important to have more transparency on the operating rules of these decisions:

- What are the operating rules? Are these rules based for example on the popularity of an account?
- How is it decided for how long a block stays in place?
- How are national laws taken into account?
- How is IP spoofing dealt with?
- When is it decided to suspend an account indefinitely?
- How is it decided when content is sensitive or in conflict with the community standards?

Transparency about online tools

We are aware that social media platforms use software to detect and remove pornographic and paedophilic content. We are not aware of any software used by them to target hate speech: racism, antisemitism, homophobia, sexism, etc. The developments regarding these kinds of software are for many reasons of interest for trusted flaggers. It will show the level of professionalism that social media uses to tackle hate speech and with that it will help in the collaboration with trusted flaggers:

- Do social media companies use software to detect and remove hate speech?
- If yes, what kind of software?
- If no, are they planning to start using such filtering algorithms in the future?

SIMPLIFICATION OF THE SOCIAL NETWORKS' REPORTING SYSTEMS

Take measures to bridge the existing gap and improve feedback procedures on reports made by trusted flaggers and regular users[3].

Simplifying reporting systems is not just important for trusted flaggers, but at least as relevant for the users of social media. Here we will focus on the general user reporting systems, which can be used by every user. Our main focus is how social media can further improve their reporting systems. When making a report it is sometimes unclear which one of the many options is the right one, or sometimes the reporting process is long, or there is simply no response from the platform. These are some of the causes that will discourage the user to continue with the report or to make a report in the future. However, we have to involve the users in the reporting process in order to remind them of their rights and obligation as citizens; social networks are our new 'streets', our 'digital streets'. As professionals we have the mission to help the users. Not just so they are able to report content, but also to do so in the proper situations.

Our technical recommendations regarding the simplification of reporting systems are based on our experience as users as well as experts.

Develop a direct access for reporting a comment

Though some platforms give quite a good overview, we have noticed some difficulties. It is often easy to report a whole message, may it be a video, photo or text, it is not always as easy to make a report on a specific comment. The URL of a specific comment is sometimes hard to find and with that, hard to search for later. This makes it harder to report it at a later hour or to a trusted flagger organization, like INACH.

[3] INACH Annual Conference 2017, Vienna, 12 October 2017, recommendations presented to the OSCE INTERNET FREEDOM CONFERENCE - THE ROLE AND RESPONSIBILITIES OF INTERNET INTERMEDIARIES held at the Hofburg in Vienna: http://www.inach.net/fileadmin/user_upload/INACH2017Final_Recommendations-OSCE.pdf

Develop the possibility to report multiple comments on a message, a page, or an account, in one report

Based on our experience, there is usually more than one hateful comment under a post, a tweet, or a video. For now, users have to report each comment separately. In order to save time, it would be easier to be able to report multiple hateful comments on one message, page or account, in one reporting.

Responsibility to your users to respond on every report made and the proposal of a dashboard

The response rate and time has improved in the last years, but unfortunately it is not perfect yet. It is not always clear on which report the platform gives their response, you do not always get a response on every report made, e.g. on a comment, and more annoyingly, sometimes a response is given without any reference number or link. Supplementing the current improvements on response rate, a clear overview or dashboard of reports would be a great improvement. Especially trusted flaggers make numerous reports, which make an overview necessary. Right now, everyone tracks their own reports and responses of the platforms. The lack of overview is not only a source of annoyance, but also gives way for errors in the numbers of responses, response times and analysis made based on these numbers. It is obvious to see that this, among other things, may lead to trusted flaggers not being able to provide data that is completely correct on the issue of cyber hate removal from social media platforms.