

# CHALLENGING DIGITAL ANTIGYPSYISM



ALBANIA | SERBIA | TURKEY | UKRAINE

CHALLENGING DISCRIMINATION PROMOTING EQUALITY



Copyright: ©European Roma Rights Centre, April 2023

Please see [www.errc.org/permissions](http://www.errc.org/permissions) for more information about using, sharing, and citing this and other ERRC materials

Authors: **Milena Ćuk and Bernard Rorke**

Contributors: **Serkan Baysak, Ebru Ertaş, Dragana Kokora, Roxhers Lufta, Aleksandar Smalović, Nataliia Tomenko, and Volodymyr Yakovenko**

Data Analyst: **Vladan Minić**

Editor: **Hannah Crane**

Cover Design: **Sophio Datishvili**

Layout: **Dzavit Berisha**

Cover photo: © **ERRC 2023**

This report is published in English.

The ERRC would like to thank all the volunteers who participated in this project. This report would not have been possible without the time and support they generously gave to monitoring, reporting, and documenting online hate speech:

**Albania:** Franko Veliu, Sara Dungaj, Kujtim Mile, Altin Xhaferri, Eleni Nanaj, Bukurie Karoshi, Emiliano Elmazi, Olsi Bilani, Elvi Nako, Eglja Aliu, Olta Karoshi, Luan Avdiu, Fevronia Gjini, and Aldo Cela.

**Serbia:** Tamara Kovačević, Elvin Ramadani, Sadik Saitović, Mangala Ljatići, Đode Ličić, and Marko Radaković.

**Turkey:** Tuğçe Tural, Dündar Kılıç, Gülgün, Yasin Polat, Tugay Aslım, and Fatoş Kaytan, Emre Balon, Ayça Tural, Tuğana Tuba Özcan, Gökten Yıldırım, Ezgi Aydın, Bilge Çiçekli, Nurdan Geçkalan, Yeşim Yolcu, Pınar Aktoy, and Osman Küçükler.

**Ukraine:** Anastasia Zuravel, Arsenij Kniazkov, Victor Chovka, Elvira Popenko, Diana Grigorichenko, and Oleg Artemchuk.

This project was funded by Swedish International Development Agency.

**Address:** Avenue de Cortenbergh 71, 4th floor, 1000 Brussels, Belgium

**E-mail:** [office@errc.org](mailto:office@errc.org)

[www.errc.org](http://www.errc.org)

## SUPPORT THE ERRC

The European Roma Rights Centre is dependent upon the generosity of individual donors for its continued existence. Please join in enabling its future with a contribution. Gifts of all sizes are welcome and can be made via PAYPAL on the ERRC website ([www.errc.org](http://www.errc.org), click on the Donate button at the top right of the home page) or bank transfer to the ERRC account:

Bank account holder: **EUROPEAN ROMA RIGHTS CENTRE**

Bank name: **KBC BRUSSELS**

IBAN: **BE70 7360 5272 5325**

SWIFT code: **KREDBEBB**

# Table of Contents

<b>Executive summary</b>	<b>4</b>
<b>Introduction</b>	<b>9</b>
About the project and the aim of the research	10
Defining hate speech: debates and dilemmas	11
Working definition of hate speech for the project	17
Defining antigypsyism	19
<b>Social networks and their codes of conduct</b>	<b>21</b>
Previous monitoring of social networks and their codes of conduct	22
<b>Common anti-Romani attitudes and narratives</b>	<b>25</b>
<b>Context of hate speech and antigypsyism in the target countries</b>	<b>28</b>
Albania	28
Serbia	29
Turkey	30
Ukraine	31
<b>Methodology</b>	<b>32</b>
Data Collection	32
Processing data	32
Qualitative data	33
Limitations of the research	34
<b>Findings</b>	<b>35</b>
Albania	35
Serbia	41
Turkey	56
Ukraine	67
<b>General Conclusions</b>	<b>75</b>

## Executive summary

This report was produced by participants of the ERRRC’s volunteer-led project *Challenging Digital Antigypsyism*. The volunteers, who formed digital activist communities, were tasked with monitoring and recording examples of hate speech targeting Roma on online media and social networks, and reporting this hate speech by using available tools on each platform. The reporting took place in four non-EU countries: Albania, Serbia, Turkey, and Ukraine, and was conducted between November 2020 and August 2021.

The volunteers were driven by concern over the prevalence of anti-Roma hate speech and a desire *to do something about it*, to develop practical and effective responses to counter online hatred and its consequences. Many of the volunteers felt that hate speech targeting Roma had been overlooked for too long. More generally, there is widening concern about the spread and consequences of unchecked hate. According to a recent European Parliament (EP) study, hate speech and hate crime ‘poison society’ and have been steadily on the rise over the last decade; hate speech has surfaced at the “*highest level of the public administration of some Member States*”. The EP report noted how hate speech has become especially prevalent on social media, where both political actors and citizens “*express their thoughts without inhibition*”, and how attempts to regulate hate speech on social media so far have brought ambiguous effect.<sup>1</sup>

In fact, as haters carry on hating, they gain increasing notoriety and influence not just across social media but also within the political sphere, posing a serious threat to values of plurality and tolerance. Beside soft measures that serve to build social resilience against hate speech, the authors of the EP report also recommend ‘hard measures’ to create a solid framework and institutional network to tackle hate speech and hate crime.<sup>2</sup>

When it comes to Roma specifically, Fernand de Varennes, the UN special rapporteur on minority issues, on the occasion of Roma Holocaust Memorial Day, called on states to “*do more to proactively combat rising signs of intolerance and attacks against Romani and other minorities, particularly hate crimes and attacks on social media.*” Recalling the fate of those who were portrayed as alien and antagonistic to the nation in Nazi Germany, de Varennes condemned the fact that “*today the Roma are again facing the same sort of divisive rhetoric*”, scapegoated by politicians and demonised and targeted on social media.<sup>3</sup> Within this context, the volunteer researchers in each of the four countries aimed to: (i) build a profile of those toxic spaces on different social networks and online media where anti-Romani speech most commonly occurs in each of the

1 European Parliament Committee on Civil Liberties, Justice and Home Affairs (LIBE Committee), *Hate speech and hate crime in the EU and the evaluation of online content regulation approaches*, European Parliament, July 2020. Available at: [https://www.europarl.europa.eu/RegData/etudes/STUD/2020/655135/IPOL\\_STU\(2020\)655135\\_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/STUD/2020/655135/IPOL_STU(2020)655135_EN.pdf).

2 *Ibid.*

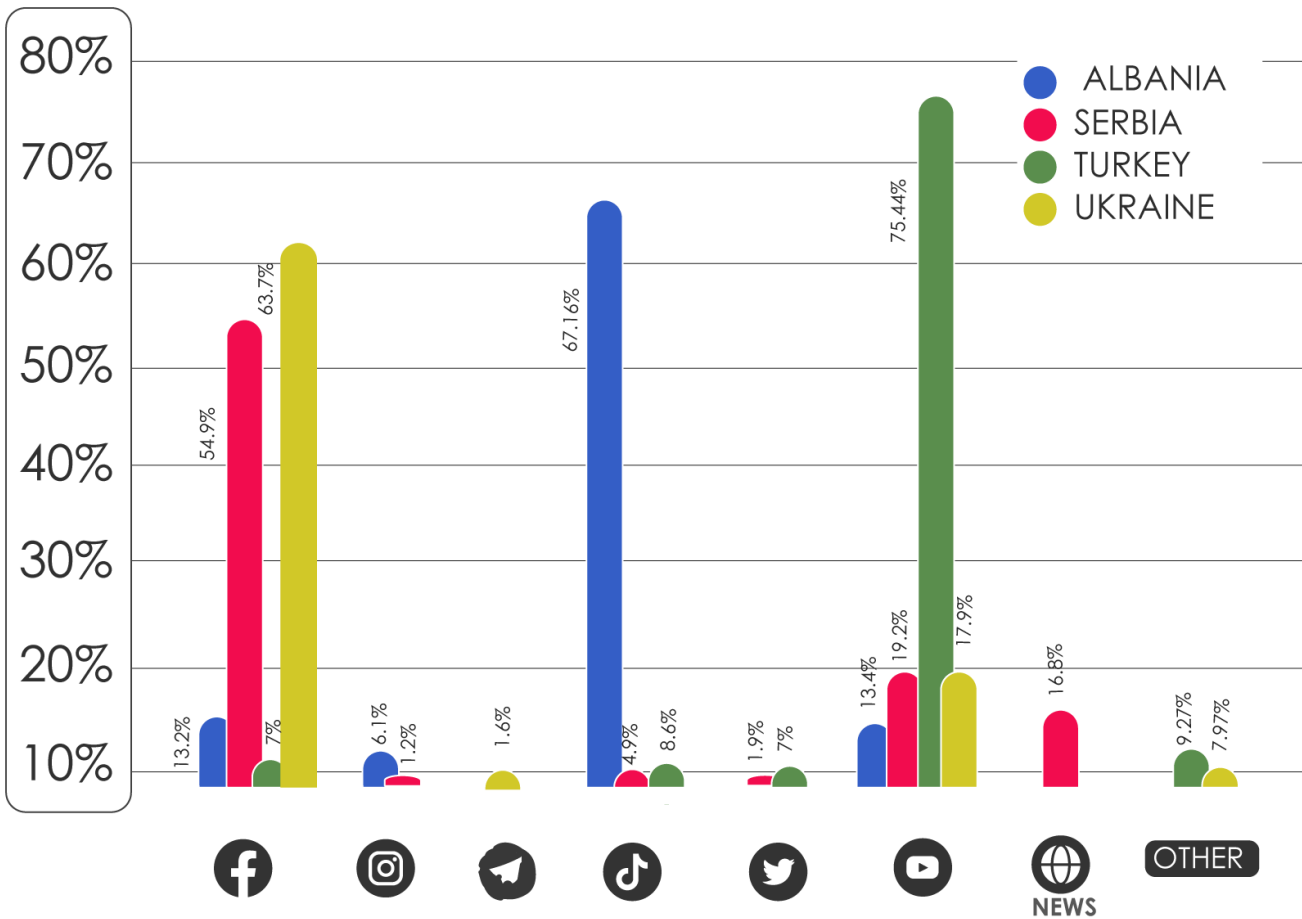
3 UN Office of the High Commissioner for Human Rights, Press Release, *UN expert urges governments to crack down on hate speech and crimes targeting Romani and minorities*, Roma Holocaust Memorial Day, 2 August 2021. Available at: <https://www.ohchr.org/en/press-releases/2021/07/un-expert-urges-governments-crack-down-hate-speech-and-crimes-targeting>.

target countries; (ii) reveal the typical tropes favoured by online haters targeting Roma; (iii) establish response and removal rates by online platforms following submission of complaints about hate content; (iv) observe to what extent platforms’ community standards were applied when it came to anti-Roma content; and (v) provide a baseline of evidence and knowledge to enable legal challenges to online hate speech that poses a danger to Roma.

As the results show, the status of much of the reported content was difficult to ascertain because some platforms don’t even have an option for reporting hate speech and, with the exception of Facebook, social networks rarely notify users about the status of the reported content. The data sets and the observations and conclusions drawn by the volunteer teams provide a vivid and representative snapshot of online anti-Roma hate speech Roma in Albania, Serbia, Turkey, and Ukraine.

**SOURCES OF HATE SPEECH CONTENT TARGETING ROMA**

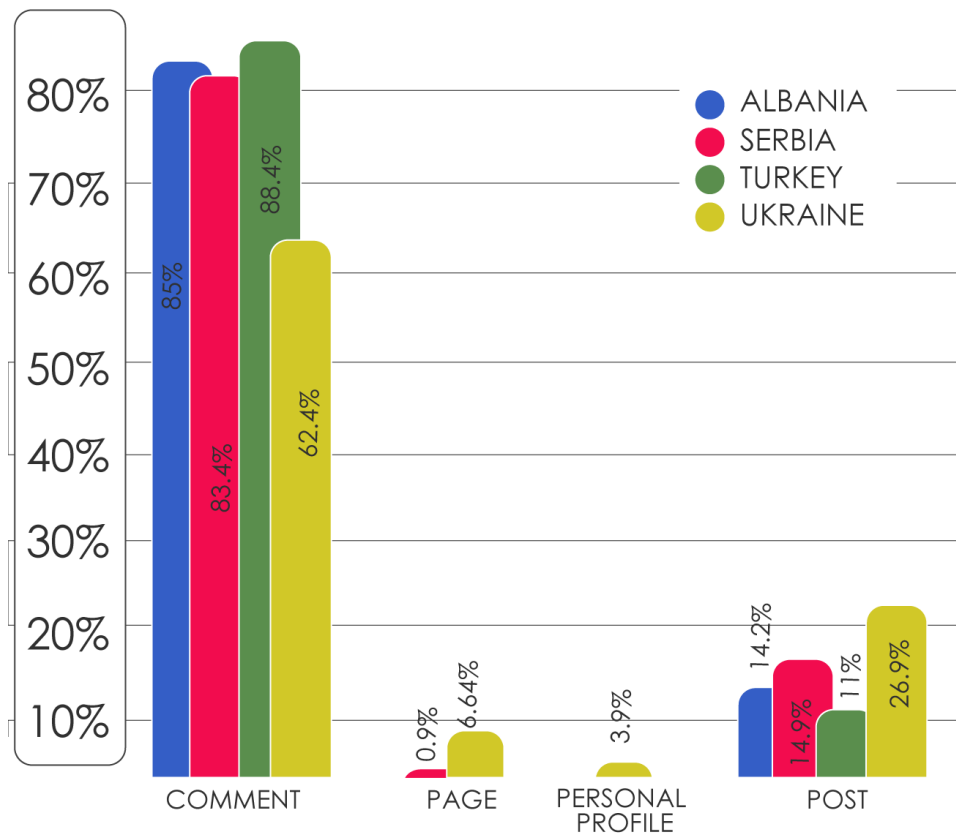
Facebook as a source was the most represented in the Serbian and Ukrainian samples, TikTok in Albania, and YouTube in Turkey. Online news portals were also a frequent source of anti-Romani content in the Serbian sample.



The most searched platforms differed between the four targeted countries due to a variety of user habits in each place and the fact that each volunteer searched networks on which they were most active.

**TYPE OF REPORTED CONTENT IN FOUR SAMPLES**

Related to the type of content, reporting of the comments was the most represented in all countries. Ukraine had the highest number of reported posts.



**PLATFORMS, HATERS, AND MODERATORS**

Facebook stood out in all four countries as the most responsive platform as regards sending notifications about the status of reported content. In stark contrast, tracking the status of reported content was most difficult at YouTube and TikTok. In Serbia, what was notable and remains critical is the lack of moderation policies concerning hate speech at online tabloid news portals.

One general conclusion was that there was a lack of consistency at Facebook and YouTube when it came to removing reported hate speech content. Volunteers also found that Facebook often fails to recognise thinly-coded hate content disparaging ‘Cig@n’ or ‘G#psies’, statements

such as ‘*Hitler knew what to do with them*’, or racist ‘dog-whistles’ targeting Roma. Dog-whistles involve strategic and coded racist manipulation, they “*trade in racist ideas, explicitly avoiding naming race directly while invoking negative racial stereotypes that the viewer often reacts to without making a conscious connection of the underlying racial division.*”<sup>4</sup> So while removal rates were better on Facebook than other platforms, and in many cases ethnic slurs were considered to be a violations of community standards, the monitors found much inconsistency, either, as mentioned above, by Facebook content moderators failing to recognise dog-whistles, or simply deciding in other reported cases that ethnic slurs did not violate community standards.

In pre-war Ukraine, volunteers identified the growing use of the messaging application Telegram by far-right groups in targeting Roma. This is part of a growing international phenomenon, worthy of closer monitoring in the future, and one that will require a response from states when it comes to the spread of online hate against Roma and other racialised minorities on Telegram.

For, as the Washington Post recently reported, Telegram plays an increasingly important role in the right-wing information ecosystem, “*offering a respite from scrutiny and moderation. It’s a place where the fringe’s bubble of disinformation and rhetoric can remain unpunctured — which is often precisely the appeal.*” As one expert described it, public Telegram channels are often used as advertisements for private channels, a feature that has made Telegram “*a space for more radical, more extreme discourses on the right.*”<sup>5</sup> Telegram’s end-to-end encrypted chat function and a lack of in-app moderation allows for uncensored communication and information sharing in oft-called ‘secret chats’.

Across Europe, Telegram has become the social media platform of choice for neo-Nazis and other extremists whose accounts on platforms such as YouTube and Facebook have already been deleted. Recently, Germany’s Federal Criminal Police (BKA) accused Telegram of consistently failing to respond to requests to delete far-right content. According to Germany’s Network Enforcement Act, it is the obligation of operators to remove illegal content. Telegram’s failure to comply has raised the possibility of it being banned in Germany.<sup>6</sup>

In the four countries, researchers found online hate speech and content disparaging Roma ranged from comedy routines where Roma are depicted as oafish, coarse, and the butt of every joke, to direct and specific neo-Nazi calls to commit acts of racist violence against Roma. Online commentary variously accuses Roma of involvement in petty theft and organised crime, welfare abuse and fraud, and being work-shy and undeserving beneficiaries of affirmative action programs. While the tone varies from sarcasm and contempt to fuming full-on race hate, there can be little doubt concerning the corrosive, cumulative impact of online hate speech which disparages and dehumanises Romani people. The research confirms that when we tackle hate speech we need to look beyond the immediate threat to public order, to the prevailing social climate in which antigypsyism has been effectively normalised. The evidence from the report highlights the need to fight both the immediate danger and change the prevailing climate.

4 Yahaira Cacereson, ‘Decoding Racist and Xenophobic Dog-Whistles’, *America’s Voice*, 9 November 2021. Available at: <https://americasvoice.org/blog/decoding-racist-and-xenophobic-dog-whistles/>.

5 Philip Bump, ‘The platform where the right-wing bubble is least likely to pop’, *Washington Post*, 23 April 2022. Available at: <https://www.washingtonpost.com/politics/2022/04/23/telegram-platform-right-wing/>.

6 DW, *German police pressure Telegram to delete far-right content*, January 17, 2022. Available at: <https://www.dw.com/en/german-police-pressure-telegram-to-delete-far-right-content/a-60453402>.

Such is the alarm about this climate of hate that on 9 December 2021 the European Commission presented an initiative<sup>7</sup> to extend the list of ‘EU crimes’ to hate speech and hate crime. This was in response to what the Commission described as the particularly serious and worrying phenomenon of the “sharp rise in hate speech and hate crime across Europe – offline and online.” Commission Vice-President for Values and Transparency, Věra Jourová, said:

*“Hate has no place in Europe. It goes against our fundamental values and principles. We need EU action to make sure that hate is criminalised the same way everywhere in Europe.”*<sup>8</sup>

7 European Commission General Publications, *A more inclusive and protective Europe: extending the list of EU crimes to hate speech and hate crime*, 26 November 2021. Available at: [https://ec.europa.eu/info/files/communication-extending-eu-crimes-hate-speech-and-hate-crime\\_en](https://ec.europa.eu/info/files/communication-extending-eu-crimes-hate-speech-and-hate-crime_en).

8 European Commission Press Release, *The Commission proposes to extend the list of ‘EU crimes’ to hate speech and hate crime*, 9 December 2021. Available at: [https://ec.europa.eu/commission/presscorner/detail/en/ip\\_21\\_6561?fbclid=IwAR1F5ebLg-11ib1tVi\\_-O7WMfRZnE4g4Ga35Ozfu1vBq-PuXTCjXDghoE8I](https://ec.europa.eu/commission/presscorner/detail/en/ip_21_6561?fbclid=IwAR1F5ebLg-11ib1tVi_-O7WMfRZnE4g4Ga35Ozfu1vBq-PuXTCjXDghoE8I).



## Introduction

When there is an emboldened section of racists inciting hatred and violence online, eventually there will be an increase in hate crimes in the real world. These things are measurably linked in many countries.

The European Roma Rights Centre (ERRC) has been illuminating the link between online hate speech targeting Roma and real-world violence in various countries for a number of years, and particularly when this intersects with far-right movements and politicians. There is a direct link between the hate speech from far-right political movements and online antigypsyism; the rise of the first emboldens the second and gives them the confidence to use hate speech online that previously was taboo in the public sphere.

This project was borne from a need by Romani activists to *do* something about the clear danger of escalation when hate speech online goes viral. ERRC activists and volunteers felt that the threat of escalation of online hate speech had often been overlooked or played down. As it is always difficult to predict what type of online hate will become hate crimes offline and how, it is important to always treat instances of online viral hate speech very seriously.



### About the project and the aim of the research

In early 2020, the ERRC established the volunteer-led project *Challenging Digital Antigypsyism*, which aimed to challenge online hate speech against Roma through the creation of digital activist communities focused on reporting and countering hate speech on social media platforms.

In the first phase, cohorts of Romani activists from Albania, Serbia, Turkey, and Ukraine were recruited as national volunteer coordinators for each country. They were trained by ERRC staff and guest experts to recognise hate speech targeting Roma and volunteer management, and then prepared for a process of monitoring selected online platforms. These National Volunteer Coordinators, with the ERRC's supervision, recruited and selected national teams of volunteers (ERRC Roma Rights Defenders) and developed these groups into semi-autonomous activist communities. The National Volunteer Coordinators used the knowledge and skills they had gained to train and manage these local groups of volunteers, who undertook activities designed to challenge the environment of racial hatred which often exists on social media platforms.

The main task of the volunteers in each country was to monitor and record examples of hate speech targeting Roma on online media and social networks, and to report it using available tools on each platform. Without wanting to be drawn too much into academic debates of how to define hate speech, a working definition was created for practicable use in this project. This definition was based largely on the community guidelines of different social media platforms and served to define content which would likely be considered hate speech by the main social media providers. Volunteers were also tasked with recording and storing evidence of reported content for later analysis.

#### The aims of the research:

- Build a profile of common spaces on different social networks and online media where anti-Romani speech most commonly occurs in each of the target countries.
- Investigate what kind of reported content is being removed from the networks after being reported by users (in relation to the project's working hate speech definition).
- Check the responsiveness (post-report communication) of different online platforms/networks after a report is submitted for violating community standards, and to observe any differences between platforms in the accuracy of implementing standards and removing content found to violate their standards.
- Uncover typical anti-Roma narratives used online in each of the target countries through patterns of content reported by volunteers.
- Compare any commonalities or marked differences in digital antigypsyism recorded in the target countries.
- Provide a baseline of evidence and knowledge for each country, from which legal challenges could be launched in relation to dangerous hate speech online against Roma.

This research constitutes a pilot study of online hate speech targeting Roma in four non-EU countries, as well as the efficiency of networks to respond to reports relating to Roma in four languages which are relatively small in terms of users when compared to languages such as

English. Lessons learned in the production of this research have served to improve research protocols for monitoring within the ongoing project. The results of this study were also intended for use in tailoring further advocacy activities related to challenging digital antigypsyism by the ERRC and its volunteer Roma Rights Defenders.

## Defining hate speech: debates and dilemmas

The concept of ‘hate speech’ remains as fraught, politically charged, and contested as ever. It is not only the fact that a universally accepted definition of hate speech is lacking, but the uncertainty over what is and what is not ‘hate speech’ adds to the confusion and disagreement over how best democracies should respond to the threats posed by those who wilfully peddle hatred.<sup>9</sup>

At the risk of oversimplification, the divide is between, on the one hand, those who oppose all restrictions on free speech, who argue that the best antidote to hate speech is more speech in the ‘market place of ideas’; and on the other, those who call for bans on hate speech on the grounds that public expression and promotion of hatred poisons relations between groups and individuals, endangering those at the receiving end. From this standpoint, the crucial question is about the direct targets of the abuse:

*“The harm that expressions of racial hatred do is harm in the first instance to the groups who are denounced or bestialized in pamphlets, billboards, talk radio and blogs ... Can their lives be led, can their children be brought up, can their hopes be maintained and their worst fears dispelled, in a social environment polluted by these materials?”<sup>10</sup>*

Those in favour of content-based bans reject ‘the market-place of ideas’ metaphor for free speech as a neo-liberal delusion that reinforces structural inequalities. They typically argue that substantive equality requires the prohibition of hate speech that targets historically excluded minority groups, to enable them to renegotiate their power and standing in society. The idea that the best response to hate speech is more speech presumes the market-place is a neutral and level play-zone, and naively imagines that false ideas will always lose in the struggle with true ones, and that good will prevail, as if ideas “operate in a social vacuum.”<sup>11</sup>

Those who oppose such bans question whether hate speech is at all a useful concept, and argue that restrictions on free speech undermine democracy itself and the legitimacy of those in power. In addition to principled objections, sceptics question the efficacy of such bans and suggest that a focus on hate speech prohibitions runs the risk of diverting political energies from more meaningful political responses to the underlying causes of racism.

In basic terms, the challenge is to reconcile Article 19 of the International Covenant on Civil and Political Rights (ICCPR) which guarantees freedom of expression, with Article 20(2)

9 Bernard Rorke, *Free to hate? Anti-Gypsyism in 21<sup>st</sup> Century Europe*, in Molnar (ed.), *Free Speech and Censorship around the Globe*. CEU Press, 2015.

10 Jeremy Waldron, ‘Free speech and the Menace of Hysteria’, *New York Review of Books*, 29 May 2008. Available at: <https://www.nybooks.com/articles/2008/05/29/free-speech-the-menace-of-hysteria/>.

11 Bhikhu Parekh, ‘Is there a case for banning hate speech?’, in *The Content and Context of Hate Speech: Rethinking Regulation and Responses*, Michael Herz and Peter Molnar (eds), pp. 37-56. Cambridge University Press, 2012.

which sets limitations on freedom of expression and requires States to “prohibit” certain forms of speech which are intended to sow hatred. States must find a way to meet their democratic obligations to protect the fundamental rights of all citizens, and counter all forms of direct and indirect discrimination while at the same time protecting freedom of expression.

As public consensus continues to shift around what constitutes discrimination, and understandings of the harms caused by inequalities evolve, so too have definitions of hate speech, as those who would regulate what legally constitutes hate adapt to new situations, or react to particular incidents or pernicious phenomena. Safe to say it’s unlikely that the answers to the vexed questions regarding the nature and the scope of governmental responses could ever be uniform across all countries, settings, and situations.

### Context matters

ARTICLE 19<sup>12</sup>, an NGO concerned about vague definitions of what constitutes hate speech and overly broad prohibitions in national laws, believes that effective and nuanced responses to ‘hate speech’ are critical, and emphasises the importance of context. In terms of incitement to violence, *“the relationship between speech and action is always a contextual matter, never a matter just of the content of the speech.”*<sup>13</sup> When it comes to antigypsyism, context matters, and in seeking to establish clear boundaries between permissible and impermissible forms of expression, ARTICLE 19 describes context in the following terms:

*“The context may be characterised by frequent acts of violence against individuals or groups based on prohibited grounds; regular and frequently negative media reports against/on particular groups; violent conflicts where groups or the police oppose other groups; reports raising levels of insecurity and unrest within the population.”*<sup>14</sup>

This is a characterisation of the ‘context’ that is familiar to many Romani citizens across Europe. In terms of context, the torrent of disparaging, discriminatory, and hateful speech targeting Roma fosters and sustains a broad and toxic consensus that ‘Gypsies get what they deserve’. Such speech dehumanises and degrades Romani people. Beyond the hard core of haters, hate speech contaminates the public sphere in a manner that inhibits any sense of solidarity or empathy. The cumulative effect is that majority populations fail to recognise discriminatory treatment of Roma for what it is: unreconstructed racism.

### The deeply damaging long-term consequence of hate speech

In his examination of the case for the banning of hate speech, Bhikhu Parekh says: *“Hate speech is a distinct kind of speech and much conceptual confusion is created – and the net of prohibition unduly widened – by subsuming all forms of uncivil and hurtful speech around it.”* Parekh provides what he terms a ‘reasonably precise meaning’, and ascribes three essential features to hate speech:

<sup>12</sup> See: <https://www.article19.org/>.

<sup>13</sup> Peter Molnar, ‘Responding to ‘Hate Speech’’, in *The Content and Context of Hate Speech: Rethinking Regulation and Responses*, Michael Herz and Peter Molnar (eds), pp. 183-197. Cambridge University Press, 2012.

<sup>14</sup> ARTICLE 19, *Prohibiting incitement to discrimination, hostility or violence*, Policy Brief, December 2012. Available at: <https://www.refworld.org/pd/fid/50bf56ee2.pdf>.

1. *that it is directed against a specified or easily identifiable group of individuals based on an arbitrary and normatively irrelevant feature;*
2. *it stigmatises the target group by ascribing to it highly undesirable qualities;*
3. *the target group is viewed as an undesirable presence and a legitimate object of hostility to be expelled, exterminated or subjected to discrimination.*<sup>15</sup>

Much of the anti-Roma hate speech common in 21st-century Europe falls squarely within the broad parameters of Parekh's three essential features. Having set out what is distinctive about hate speech to prevent the net of prohibition being 'unduly widened,' Parekh then proceeds to dismiss notions of imminent danger, stretching the net of prohibition to include speech which may not result in violence. Parekh argues that content should be judged by its long-term effects on a targeted group rather than its immediate consequences, because:

*"If anything can be said about a group of persons with impunity, anything can also be done to it. This is because if a group can be treated with contempt, stripped of dignity, dehumanised, treated as belonging to an inferior species, and a moral climate is created in which harm done to it is seen as right and proper and does not arouse a sense of outrage."*<sup>16</sup>

Europe's Roma have been treated with contempt, dehumanised, and harm is done to them that does not arouse a sense of outrage. There is no doubt that disparaging, inflammatory, and hateful anti-Roma speech has coarsened public sensibility and strikes at the core of notions of shared belonging in a democratic polity. Parekh warns that we lose sight of the deeply damaging long-term consequence of hate speech if we only concentrate on and judge it in terms of the likely immediate threat to public order. He argues that imminent danger occurs against, and is imminent because of, the prevailing social climate, and "*consistency demands that we concentrate our efforts not only on fighting the immediate source of danger, but also on changing the climate*".<sup>17</sup>

### Free speech, hate speech, and the case against banning

In the opposing corner, Kenan Malik believes that *no* speech should be banned solely because of its content. He distinguishes 'content-based' regulation from 'effects-based' regulation and would permit the prohibition only of speech that creates imminent danger. He opposes content-based bans both as a matter of principle and with a mind to the practical impact of such bans. In principle, Malik holds that free speech for everyone except bigots is not free speech at all. The right to free speech "*only has political bite when we are forced to defend the rights of people with whose views we profoundly disagree*".

In practice, Malik asserts that you cannot reduce or eliminate bigotry simply by banning it: "*hate speech restriction is a means not of tackling bigotry but of rebranding certain, often obnoxious ideas or arguments as immoral. It is a way of making certain ideas illegitimate without bothering to challenge them. And that is dangerous.*"<sup>18</sup> Malik cites Britain as an example. In 1965, Britain prohibited

15 Bhikhu Parekh, 'Is there a case for banning hate speech?' in *The Content and Context of Hate Speech: Rethinking Regulation and Responses*, Michael Herz and Peter Molnar (eds), pp. 37-56. Cambridge University Press, 2012.

16 *Ibid.* p. 44.

17 *Ibid.* p. 46.

18 Peter Molnar, 'Interview with Kenan Malik', in *The Content and Context of Hate Speech: Rethinking Regulation and Responses*, Michael Herz and Peter Molnar (eds), pp. 81-91. Cambridge University Press, 2012.

incitement to racial hatred as part of the Race Relations Act; Malik describes the decade that followed as probably the most racist in British history:

*“It was the decade of ‘Paki-bashing’, when racist thugs would seek out Asians to beat up. It was the decade of fire-bombings, stabbings and murders. In the 1980s, I was organising street patrols in East London to protect Asian families from racist attacks. Nor were thugs the only problem. Racism was woven into the fabric of public institutions.”<sup>19</sup>*

Malik argues that Britain is a very different place today. Not that racism has disappeared, but rather that the ‘open, vicious, visceral bigotry’ that scarred the Britain he grew up in has largely ebbed away; not because of laws banning expressions of racial hatred, but rather because of broader social changes and because minorities themselves stood up and fought back.

Malik insists on the distinction between word and deed, and states that racist speech should be a moral issue and not a legal one. The exception is those circumstances where there is both a direct link between speech and violent action, and intent on the part of the speaker for that violence to be carried out. Such incitement, according to Malik, should be illegal but it has to be tightly defined.

It is precisely this consideration that prompted ARTICLE 19 to provide a clear definition of the circumstances in which certain types of hate speech can or must be limited – measures that should be used only exceptionally and as a last resort – with a view to ensuring that all people are able to enjoy both the right to freedom of expression and the right to equality.

### **ARTICLE 19’s severity threshold**

ARTICLE 19’s position is that it argues for states to engage in a range of law and policy measures to counter hate speech with more speech, seeking to maximise inclusivity, diversity, and pluralism in public discourse.

They maintain that an overly broad, all-encompassing concept of hate speech that includes any expression of discriminatory hate towards people is; *“too vague for use in identifying expression that may legitimately be restricted under international human rights law.”<sup>20</sup>*

ARTICLE 19 proposes a typology for identifying hate speech according to the severity of the expression and its impact, informed by international human rights law. This six-part comprehensive test can be used to ascertain in which situations the danger of violence, hostility, or discrimination is sufficiently present to justify prohibitions on the expression. It consists of the following criteria<sup>21</sup>:

<sup>19</sup> *Ibid.* p. 84.

<sup>20</sup> ARTICLE 19, ‘Hate speech’ explained: A summary. 7 December 2020. Available at: <https://www.article19.org/resources/hate-speech-explained-a-summary/>.

<sup>21</sup> ARTICLE19, *Prohibiting incitement to discrimination, hostility or violence Policy Brief*, December 2012 pp.29-40. Available at: <https://www.refworld.org/pd/50bf56ee2.pdf>.

## 1. CONTEXT OF THE EXPRESSION:

contextual analysis should take into account factors such as the existence of conflict in society, including recent incidents of violence against the targeted group; the existence and history of institutionalised discrimination; the legal framework, including the recognition of the targeted group's protected characteristic in any anti-discrimination provisions or lack thereof; the media landscape, for example regular and negative media reports about the targeted group with a lack of alternative sources of information; and the targeted group's situation and representation in formal political processes.

## 2. THE SPEAKER:

the position of the speaker, and their authority or influence over their audience is crucial. This analysis should also examine the relationship of the audience to the speaker, and issues such as the degree of vulnerability and fear among the various communities, including those targeted by the speaker, or whether the audience has high levels of respect or obedience of authority voices.

## 3. INTENT:

there must be (i) intent to engage in advocacy to hatred; (ii) intent to target a group on the basis of a protected characteristic, and (iii) having knowledge of the consequences of their action and knowing that the consequences will occur or might occur in the ordinary course of events (i.e. in which no unforeseeable change or event has occurred).

## 4. CONTENT OF THE EXPRESSION:

the audience's understanding of the content is particularly important, in particular where the expression contained direct or indirect calls for discrimination, hostility or violence. International standards have recognised that certain forms of expression provide "little scope for restrictions", in particular artistic expression, public interest discourse, academic discourse and research, statements of facts and value judgements.

## 5. EXTENT AND MAGNITUDE OF THE EXPRESSION:

the analysis should examine the public nature of the expression, the means of the expression and the intensity or magnitude of the expression in terms of its frequency or volume. If the expression was disseminated through the mass media, consideration should be given to media freedom, in compliance with international standards.

## 6. LIKELIHOOD OF HARM OCCURRING, INCLUDING ITS IMMINENCE::

the probability of the harm advocated by the speaker occurring must also be established in order to measure the level of severity. Some degree of risk of resulting harm or causality must be identified; it should be demonstrated that the communication will gain some credence with the attendant result of discrimination, hostility or even violence, against the targeted group; and that the possibility of harm should be imminent – the immediacy with which the acts (discrimination, hostility or violence) called for by the speech are intended to be committed should be deemed relevant.

## Impact

The six-point test provides a guide to guard against arbitrary abuse of the power to restrict freedom of expression. This is important when it comes to Roma in Central and Eastern Europe, as there is, as the civil-libertarian Nadine Strossen notes, something of a conundrum in that those who call for more restriction, more banning of hate speech, are in fact calling for more discretionary powers to be handed to states and societies they hold to be inherently racist and discriminatory, and giving government enormous powers to suppress legitimate political discourse.

The very practical and political question is; who would sensibly cede even more powers to government to circumscribe freedom of expression in a state such as Hungary, where the ruling party has captured key institutions, weakened checks, and dispensed with balances? Who of reasoned mind would trust a government whose leader has been widely condemned as a racist to legislate on matters pertaining to hate speech,<sup>22</sup> and where the Fourth Amendment to the Fundamental Law of Hungary states that the right to freedom of speech “*may not be exercised with the aim of violating the dignity of the Hungarian nation*”?

Another practical objection is that not only does the fixation with content-based speech bans serve as a distraction from the real task of combating racial discrimination, but that suppressing expression just drives it underground to fester.

Bhikhu Parekh counters that if banning hate speech drives racists underground, so much the better, that is where they belong. As to the notion that a ban on hate speech can become an end in itself and an excuse to avoid well-conceived antidiscrimination policies, Parekh says “*this can happen, but there is no obvious reason why it should.*” He cites the examples of Britain, the Netherlands, and Australia as showing that bans on hate speech have gone hand-in-hand with wider campaigns to address the causes of racism, sexism, and homophobia by pressing for policy strategies to tackle racism and disadvantage.

Further, Parekh asserts that a legal prohibition is valuable in sending a message that the state values all members of society equally; in laying down norms of civility and clearly delineating what is and what is not acceptable in talking about or treating other members of society; and in affirming and enforcing these values, the law “*has a symbolic and educational significance, and helps shape the collective ethos.*”<sup>23</sup>

When it comes to combating antigypsyism, to countering the words, deeds, and institutional practices that denigrate and dehumanise Romani citizens in the 21st Century, it’s the practical impact that counts. Efforts to combat the consequences of incitement must be part of comprehensive state policies to promote equality and challenge all forms of racist exclusion. When governments are the culprits and the rule of law is thus endangered, the European Commission and the Council of Europe are duty-bound to intervene. For its part, civil society and civil rights defenders must remain alert to the predilection of governments to abuse powers at their disposal to suppress freedom of expression.

While the question of whether hate speech should be prohibited by law or merely discouraged by moral and social pressure remains contested, there is remarkable consensus that the law must be a measure of last resort. Most good-faith protagonists in the free speech debate agree that the state must create an enabling environment for the rights to freedom of expression and equality and non-discrimination to be realised by all; and that the most effective way to combat hate speech is through social action, public campaigning that holds authorities to account, vigorous collective action to counter racist threats, and broader social mobilisation in support of and solidarity with targeted communities. There is a growing need for such solidarity in times

22 RFL/RE, *European Parliament Leaders Condemn Orbán for ‘Openly Racist’ Remarks*, 30 July 2022. Available at: <https://www.rferl.org/a/european-parliament-orban-racist-remarks/31966820.html>.

23 Bhikhu Parekh, ‘Is there a case for banning hate speech?’ in *The Content and Context of Hate Speech: Rethinking Regulation and Responses*, Michael Herz and Peter Molnar (eds), p. 46. Cambridge University Press, 2012.



when hate's political harvest remains bountiful, and minorities are too often at the receiving end of racist abuse. Addressing the hurt and harm done to those directly targeted must remain at the heart of any effective strategy to counter the structures that reproduce racist injustice as well as combating the words and deeds of those 'who love to hate.'

## Working definition of hate speech for the project

Debate around hate speech aside, it became necessary for the purpose of this project to create a functional, working definition of hate speech as it appears on social media networks which were being monitored. For the most part, platforms which host social media content are subject only to the rules and community standards which they implement for themselves. For this reason, a simple definition alongside a checklist was created for ease of use by volunteers engaged in the project:

***Hate Speech is:** Any speech made in public, which incites hatred or discrimination, uses dehumanising language or statements of inferiority, taboo ethnic slurs, or encourages violence towards a person or group of people based on protected characteristics (i.e. ethnicity, sexual orientation, gender, religion, ability, skin colour).*

### THE CHECKLIST:

- Is it public? [MUST BE PRESENT]
- Is it about a protected group? [MUST BE PRESENT]
- Does it use dehumanizing or inferior language? [+ANY OF THESE]
- Does it encourage others to commit violence? [+ANY OF THESE]
- Does it use an ethnic slur which it's unacceptable to say in public? [+ANY OF THESE]
- Does it incite hatred or discrimination? [+ANY OF THESE]
- Does it matter who is saying it?

This checklist lists the criteria for the working definition of hate speech which volunteers used as a simple guide to help them identify cases of hate speech online. This was further expanded on during training with volunteers as follows:

#### Prerequisites for hate speech:

*Is it public? Is it about a protected group?*

These two criteria are essential prerequisites for content to be considered hate speech. If the content is posted on social media or another web portal, then it would be considered public (unless it is in a private chat or a closed group with only a small number of users). If it is about Roma, then it is about a protected group.

Hateful content as defined by social media platforms:

*Does it use dehumanising language?*

Examples of this include comparing people to animals or insects which are considered disgusting or inferior in some way (rats, cockroaches, parasites, monkeys). This can also be language which removes the “personhood” of the target – for example calling Roma “*asocials*”, “*inadaptables*”, a “*species*”.

*Does it incite hatred or discrimination?*

The subjectiveness of this question was discussed and debated with volunteers during training. This was the most common criteria across social networks’ content standards, but which had the least comprehensive definition. In practice, this question applies if the user believes that the tone, the language, or the intent of the speech will have the effect of making the audience hate Roma or engage in discrimination against Roma. This is where the boundary between racist speech and hate speech is often found and where subjectivity in the moderation of comments comes into play. Volunteers felt that intent has a big part to play in this question – if there is something which is written out of ignorance which uses racist stereotypes, it was perceived as different to content which weaponises racist stereotypes in order to radicalise others into hating Roma.

*Does it encourage others to commit violence?*

From a legal perspective this is the cleanest criteria for defining hate speech. Any speech which makes direct or indirect reference to violence against Roma violates community standards for all social media networks. The platform content standards provide no further definition of what constitutes ‘imminent threat’ in relation to encouragement to violence on their platforms.

*Does it use an ethnic slur which is unacceptable to say in public?*

What is considered a slur, and who can use such words, varies by country and by group. Many words which are considered slurs for Romani people are widely used in society in some of the target countries. However, volunteers agreed that antigypsyism in wider society cannot be an excuse, nor the benchmark, for whether to report ethnic slurs against Roma on social media or not.

*Does it matter who is saying it?*

In the training this was discussed in debates around the nature of hate speech. Although this is not included in any of the platforms content standards it was included in the checklist because of the risk of harm associated with certain profiles using hate speech online. There is a difference between an anonymous user who only has a handful of people in their audience, and someone who has many followers and a lot of influence online. The use of speech which may incite hatred causes greater harm when it comes from an influential profile which can reach many users.

Alongside the content guidelines published by the social networks Facebook,<sup>24</sup> Twitter,<sup>25</sup> and Instagram,<sup>26</sup> the working definition and checklist was formed in consultation with existing works on the topic by Article 19,<sup>27</sup> the Dangerous Speech Project,<sup>28</sup> Council of Europe materials,<sup>29</sup> and an article by Facebook’s Vice President EMEA of Public Policy.<sup>30</sup>

## Defining antigypsyism

Antigypsyism is a term that refers to the specific form of racism that targets Romani people, seen from society’s perspective as ‘Gypsies’. This project uses the term in reference to the accepted definitions of antigypsyism used by civil society organisations and international institutions. Alternative, but approximately analogous, terms which are used are anti-Roma Racism and Romaphobia.

The European Commission against Racism and Intolerance (ECRI) defines antigypsyism as:

*“a specific form of racism, an ideology founded on racial superiority, a form of dehumanisation and institutional racism nurtured by historical discrimination, which is expressed, among others, by violence, hate speech, exploitation, stigmatization and the most blatant kind of discrimination.”<sup>31</sup>;*

Ismael Cortes Gomez and Markus End give additional perspectives on the development of the term in scholarly and political discourse, describing antigypsyism as:

*“a historically constructed, persistent complex of customary racism against social groups identified under the stigma ‘gypsy’ or other related terms, and incorporates: 1. a homogenizing and essentializing perception and description of these groups; 2. the attribution of specific characteristics to them; 3.*

24 Facebook Community Standards: Hate Speech, 18 November 2020. Available at: [https://www.facebook.com/communitystandards/hate\\_speech/](https://www.facebook.com/communitystandards/hate_speech/) (updated since, changes viewable in the changelog).

25 Twitter Hateful Conduct Policy, 2020. Available at: <https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy>.

26 Instagram Community Guidelines, 2020. Available at: <https://help.instagram.com/477434105621119/>.

27 Article 19, *‘Hate Speech’ Explained – A Toolkit*, 2015. Available at: [https://www.article19.org/data/files/medialibrary/38231/‘Hate-Speech’-Explained---A-Toolkit-\(2015-Edition\).pdf](https://www.article19.org/data/files/medialibrary/38231/‘Hate-Speech’-Explained---A-Toolkit-(2015-Edition).pdf). Article 19, *The Camden Principles on Freedom of Expression and Equality*, 2009. Available at: <https://www.article19.org/data/files/pdfs/standards/the-camden-principles-on-freedom-of-expression-and-equality.pdf>.

28 Dangerous Speech Project, *Dangerous Speech: A Practical Guide*, 2021. Available at: <https://dangerousspeech.org/guide/>.

29 Council of Europe, *Guide to Human Rights for Internet Users, Recommendation CM/Rec(2014)6 and explanatory memorandum*, 2014. Available at: <https://rm.coe.int/16804d5b31>. Anne Weber, *Manual on hate speech*, Council of Europe Publishing, 2009. Available at: [https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDC\\_TMContent?documentId=0900001680665b3f](https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDC_TMContent?documentId=0900001680665b3f).

30 Richard Allan, *Hard Questions: Who Should Decide What Is Hate Speech in an Online Global Community?*, 2017. Available at: <https://about.fb.com/news/2017/06/hard-questions-hate-speech/>.

31 European Commission against Racism and intolerance (ECRI), *General Policy Recommendation No.13 on Combating Antigypsyism and Discrimination Against Roma*, 2011. Available at: <https://rm.coe.int/ecri-general-policy-recommendation-no-13-on-combating-anti-gypsyism-an/16808b5ace>.

*discriminating social structures and violent practices that emerge against that background, which have a degrading and ostracizing effect and which reproduce structural disadvantages.”<sup>32</sup>*

In October 2020, the International Holocaust Remembrance Alliance (IHRA) adopted the following non-legally binding working definition of antigypsyism/anti-Roma discrimination:

*“Antigypsyism/anti-Roma discrimination is a manifestation of individual expressions and acts as well as institutional policies and practices of marginalization, exclusion, physical violence, devaluation of Roma cultures and lifestyles, and hate speech directed at Roma as well as other individuals and groups perceived, stigmatized, or persecuted during the Nazi era, and still today, as “Gypsies.” This leads to the treatment of Roma as an alleged alien group and associates them with a series of pejorative stereotypes and distorted images that represent a specific form of racism.”<sup>33</sup>*

As an explanatory note they add that:

*“The word ‘Roma’ is used as an umbrella term which includes different related groups, whether sedentary or not, such as Roma, Travellers, Gens du voyage, Resandefolket/De resande, Sinti, Camminanti, Manouches, Kalés, Romanichels, Boyash/Rudari, Ashkalis, Égyptiens, Yéniches, Doms, Loms and Abdal that may be diverse in culture and lifestyles.”<sup>34</sup>*

<sup>32</sup> Cortez Gomez, Ismail., and End Markus, *Dimensions of Antigypsyism in Europe*, 2019. Edited by Ismail Cortez Gomez and Markus End. Brussels: European Network Against Racism (ENAR) and the Central Council of German Sinti and Roma.

<sup>33</sup> IHRA, *What is antigypsyism/anti-Roma discrimination?*, 2020. Available at: <https://www.holocaustremembrance.com/resources/working-definitions-charters/working-definition-antigypsyism-anti-roma-discrimination>.

<sup>34</sup> *Ibid.*

## Social networks and their codes of conduct

Social networks have defined terms of use for content posted on their platforms, known variably as community guidelines, community standards, conduct policies etc. Across the four target countries, Facebook and YouTube were the networks monitored more frequently by volunteers, therefore excerpts of their respective policies are presented here.

Facebook undergoes a regular review of its Community Standards policies and publishes the changes on its website. Facebook does not allow hate speech, currently defined as:



*... a direct attack against people — rather than concepts or institutions — on the basis of what we call protected characteristics: race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity and serious disease. We define attacks as violent or dehumanizing speech, harmful stereotypes, statements of inferiority, expressions of contempt, disgust or dismissal, cursing and calls for exclusion or segregation. We also prohibit the use of harmful stereotypes, which we define as dehumanizing comparisons that have historically been used to attack, intimidate, or exclude specific groups, and that are often linked with offline violence.<sup>35</sup>*

Hate speech is also expressly forbidden on YouTube. In their hate speech policy, it is noted:



*We consider content hate speech when it incites hatred or violence against groups based on protected attributes such as age, gender, race, caste, religion, sexual orientation, or veteran status. This policy also includes common forms of online hate such as dehumanizing members of these groups; characterizing them as inherently inferior or ill; promoting hateful ideology like Nazism...<sup>36</sup>*

Both platforms allow users to report content which they believe violates these policies. On Facebook the user launches a review of the offending content when they report it. This content is analysed by a combination of human moderators and machine learning AI to determine if it breaches the community standards.<sup>37</sup> The user is later sent a notification about their review decision (defined under a category post-report communication). If Facebook finds that their community standards have been violated, they either remove the content altogether or display a warning message when other users view this content.<sup>38</sup> There have been several former Facebook moderator whistle-blowers who have publicly criticised Facebook

35 Facebook Community Standards: Hate Speech, 2022. Available at: [https://www.facebook.com/community-standards/hate\\_speech/](https://www.facebook.com/community-standards/hate_speech/).

36 YouTube, *How does YouTube protect the community from hate and harassment?*, 2022. Available at: <https://www.youtube.com/howyoutubeworks/our-commitments/standing-up-to-hate/>.

37 Gadgets 360, *Facebook Is Now Using AI to Help Human Moderators Identify Posts That Need Review*, 2020. Available at: <https://gadgets.ndtv.com/social-networking/news/facebook-content-moderation-ai-machine-learning-update-rights-manager-2326393>.

38 Facebook Community standards: <https://transparency.fb.com/policies/community-standards/> (accessed 7 December 2021).

for workers' rights issues, including poor safeguarding policies, lack of psychological support for moderators, and intimidation in the workplace.<sup>39</sup> All of which have led to a less than perfect moderation system for flagged content on the platform.

YouTube also enforces its content policies using a combination of human reviewers and machine learning. In addition, they have developed the YouTube Trusted Flagger programme to provide robust content reporting processes to non-governmental organisations (NGOs) with expertise in a policy area, government agencies, and individuals with high flagging accuracy rates. YouTube prioritise these reports for review.<sup>40</sup> If it is established that reported content violates their policies, the content is removed and a notice sent to the content creator. There is a scale of possible further actions which can be taken, from a warning notice to the termination of the channel that violated the content policies.<sup>41</sup> Unlike Facebook, YouTube does not notify a user regarding the status of a report made about a comment underneath content. This makes it less accountable as it is more difficult to know whether reports have been successful or not.

### Previous monitoring of social networks and their codes of conduct

The most systematic and relevant data on the effectiveness of social networks to act reported content is found in the monitoring reports related to the 'Code of conduct on countering illegal hate speech online', agreed between European Commission and Facebook, Microsoft, Twitter, and YouTube<sup>42</sup> in May 2016. Instagram later joined this agreement in 2018 and TikTok in 2020.<sup>43</sup>

This code of conduct is based on the Framework Decision 2008/913/JHA of 28<sup>th</sup> November 2008 on combating certain forms and expressions of racism and xenophobia. This defines illegal hate speech as "*...all conduct publicly inciting to violence or hatred directed against a group of persons or a member of such a group defined by reference to race, colour, religion, descent or national or ethnic origin.*"<sup>44</sup>

Some of the commitments agreed by the IT companies in the code of conduct are: to have clearly defined rules (Community Guidelines) about prohibition of the promotion of violence and hateful conduct, including effective procedures to review notifications related to illegal hate speech on their platform; to have trained and dedicated team reviewing notifications (submitted reports); to review the majority of valid notifications in less than 24 hours, and to remove or disable access to reported content if necessary.<sup>45</sup>

39 BBC Facebook moderator, 'Every day was a nightmare', 2021. Available at: <https://www.bbc.com/news/technology-57088382>.

40 YouTube Trusted Flagger program: <https://support.google.com/youtube/answer/7554338?hl=en> (accessed December 2021).

41 YouTube Hate speech policy: <https://support.google.com/youtube/answer/2801939>.

42 Referred to also as IT companies in the document.

43 European Commission, *The EU Code of conduct on countering illegal hate speech online*. Available at: [https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combating-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online\\_en](https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combating-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en) (accessed 23 November 2021).

44 European Commission, *Code of Conduct on Countering Illegal Hate Speech Online*, 1 December 2016.

45 *Ibid.*

The implementation of these commitments is monitored through official and shadow monitoring. The same reporting and monitoring methodology is used in both types of monitoring so that results can be compared. Under this monitoring only EU member states are included, and similar monitoring exercises are not found in relation to the four target countries covered within this report.

Shadow monitoring of illegal hate speech on social media platforms conducted between 20th of January and 29th of February 2020<sup>46</sup> (in Romania, Slovakia, Estonia, Poland, and Spain) found the overall responsiveness and removal rate to be low. Facebook was the most responsive and efficient platform while Twitter and YouTube were found to be the least responsive platforms (TikTok was not monitored in this study).

TikTok was included in monitoring for the first time in the 6<sup>th</sup> evaluation of the Code of Conduct, in 2021.<sup>47</sup> This was the latest official/regular evaluation in which 35 organisations from 22 Member States participated. During a period of approximately six weeks (1 March to 14 April 2021), a total of 4543 notifications were submitted to the IT companies relating to hate speech which was deemed illegal (3237 notifications submitted through the reporting channels available to general users, and 1306 submitted through specific channels available only to trusted flaggers/reporters).

Results during this evaluation period show that Facebook removed 70.2% of the reported content, Instagram 66.2%, YouTube 58.8%, and Twitter 49.8%. Twitter and Instagram made progress compared to 2020, while Facebook and YouTube had higher removal rates during the previous monitoring period in 2020. TikTok had a good first result, with 80.1% removals.<sup>48</sup>

The Code of Conduct advises that the majority of notifications are assessed within 24 hours. Facebook assessed notifications in less than 24 hours in 81.5% of the cases and an additional 10.6% in less than 48 hours. The corresponding results for YouTube are 88.8% and 6.7% ; for TikTok 82.5% and 9.7%; for Twitter 81.8% and 8.9%; and for Instagram 62.4% and 17.6%, respectively.<sup>49</sup>

Facebook is found to be the most systematic and efficient in providing feedback to notifications (informing users about the status of the reported content) while YouTube is the worst in this sense. Facebook gave feedback to 86.9% of the notifications; Twitter to 54.1%; Instagram to 41.9%; TikTok sent feedback to 28.7%; and YouTube only to 7.3%. While Facebook is the only company informing consistently both trusted flaggers and general users, other IT companies provide feedback more frequently when notifications come from trusted flaggers.<sup>50</sup>

However, during shadow monitoring<sup>51</sup> that took place after the official period of monitoring (covered by the 6<sup>th</sup> evaluation of Code of Conduct) it was found that removal rates on Instagram

46 OpCode. *Monitoring and Reporting Illegal Hate Speech: Shadow Monitoring Report – first edition*, 2020.

47 European Commission, *Countering illegal hate speech online: 6th evaluation of the Code of Conduct*. Prepared by Didier Reynders, Commissioner for Justice. Factsheet, 7 October 2021.

48 *Ibid.*

49 *Ibid.*

50 *Ibid.*

51 INACH, *First Shadow Monitoring Report 2021*. Compiled by Maia Feijoo and Tamás Berecz, 2021.

and YouTube decreased by around 10% while the removal rate on Facebook was almost exactly the same as in the official period. Related to feedback rate - meaning to what extent do companies provide clear and timely feedback to reports - Facebook shows the best results with 85.6% during official monitoring. Still, during shadow monitoring, this rate decreased to 65.9%. Feedback provided by YouTube is zero in both official and shadow monitoring.

This shadow reporting shed light on the high discrepancies between the removal rates in different countries. For instance, the removal rate in Lithuania during shadow reporting was 92.7% and in Portugal only 28.6%. This indicates that monitoring should be country related and asks questions of the estimations of removals of hate speech by the social networks.

Similar monitoring studies were not found for non-EU countries.



## Common anti-Romani attitudes and narratives

There is empirical proof that both negative stereotypes and negative prejudices correlate with discriminatory behaviour.<sup>52</sup> Stereotypes are a basis for racial or other prejudice. Moreover, stereotypes can serve as a justification to prejudiced people for hostility and the negative feelings they have toward a particular group.<sup>53</sup>

Throughout Europe a persistent, negative image of Roma exists in the majority of societies.<sup>54</sup> Common racist rhetoric surrounding Roma includes the belief that they are unclean; that they are thieves; and that they cannot be trusted.<sup>55</sup> Racist stereotypes that depict Roma as lazy and irresponsible are more and more frequent.<sup>56</sup> A widely believed inaccuracy is that a major part of a given country's social expenditure goes to Romani families at the expense of other vulnerable members of the majority population, a belief which increases hostility towards Roma.<sup>57</sup>

Concerning education, a common refrain is that Romani families are 'not interested in education' as a supposition to why parents are sometimes reluctant to put their children in the hands of educators.<sup>58</sup> This is an argument that implies that Romani communities and cultures are unchanging.<sup>59</sup>

Examining the perception of Roma in Turkey, it was found that Roma are mostly associated with the word "theft" In another study, 'criminal', 'immoral', 'fickle' were found to be dominant negative perceptions of Roma by the majority of Turkish people<sup>60</sup>. Authors of this study argue that these perceptions are particularly reinforced by ideological reproductions and repetitions in the media. What is particularly worrying is that preservation of the biased image of Roma is supported by the state's institutions, which is illustrated in research by Ali Rafet Özkan, funded

52 Dovidio, John F. et al. 'Stereotyping, Prejudice, and Discrimination: Another Look', Chap. 9 in *Stereotypes & Stereotyping*, Edited by C. Neil Macrae, Charles Stangor, and Miles Hewstone. New York, The Guilford Press, 1997, pp. 278 – 311.

53 *Ibid.* p. 292.

54 Council of Europe, *Roma and Travellers; Documents Defending Roma Human Rights in Europe*. Available at: [http://www.coe.int/t/dg3/romatravellers/source/documents/defendingRomarights\\_en.pdf](http://www.coe.int/t/dg3/romatravellers/source/documents/defendingRomarights_en.pdf) (accessed 20 September 2011).

55 Henry Scicluna, *Anti-Romani Speech in Europe's Public Space - The Mechanism of Hate Speech*, 21 November 2007. Available at: <http://www.errc.org/cikk.php?cikk=2912> (accessed 15 September 2011).

56 Csepeli, György, and Dávid Simon, 'Construction of Roma Identity in Eastern and Central Europe: Perception and Self-identification', *Journal of Ethnic & Migration Studies* 30, no. 1, January 2004, pp. 129-150. *Academic Search Complete*, EBSCO host (accessed September 21, 2011).

57 *Ibid.*

58 Save the Children, *Denied a Future? The Right to Education of Roma/Gipsy and Traveller Children, Volume four – Summary*, Save the Children, 2001, p 29. Available at: [http://www.savethechildren.org.uk/en/docs/denied\\_future\\_summ.pdf](http://www.savethechildren.org.uk/en/docs/denied_future_summ.pdf) (accessed 15 April 2011).

59 *Ibid.*

60 Uştuk, Ozan, and Ayça Tunç Cox, 'Roma People of Turkey Re-Write Their Cinematographic Images', *Ethnicities* 20, no. 3, June 2020, pp. 501–19. Available at: <https://doi.org/10.1177/1468796819890463> (accessed November 2021).

by the Turkish Ministry of Culture, where Roma are depicted as “dirty, primitive, socially and culturally despicable, illiterate, polygamous and promiscuous”.<sup>61</sup>

Ismail Cortes claims that antigypsyist hate speech constitutes a core mechanism of racialisation directed towards Roma, which results in a concrete form of symbolic violence and racial discrimination that must be combated by institutional means.<sup>62</sup> Cortes analysed illustrative cases of antigypsyist hate speech in the context of the COVID-19 pandemic in Spain, and concluded a common denominator to be “...the presumption of fundamental moral differences between “them” and “us” (bad and good citizens); which symbolically (re)activated inherited group divisions among Roma and non-Roma.”<sup>63</sup>

According to a comparative report on the phenomena of online antigypsyism, it was concluded that “...antigypsyism online is on the rise and there is a continuous trend of normalization of hate speech against Roma.”<sup>64</sup> Simultaneously the trend of Roma Holocaust denial continues, in some cases by politicians.<sup>65</sup>

In the sixth round of monitoring of the Code of conduct on countering illegal hate speech online, antigypsyism was present in 12.5% of reported cases, just after sexual orientation (18.2%) and xenophobia including anti-migrant hatred (18%) as the most commonly reported grounds of hatred.<sup>66</sup> It shows that online hatred targeting Roma is highly present in European Union countries.

In the shadow report from 2020 on the implementation of the Code of conduct, it has been shown that there are differences in the most frequent hate content in targeted countries. Anti-Romani racism was found most frequently in Romania and Slovakia, while in other countries hate speech was usually found in the context of homophobia, and xenophobia including anti-refugee hatred and/or antisemitism.<sup>67</sup>

“Fake news, hoaxes and manipulations are being used as instruments for spreading hatred. The trend of so-called humorous racism is also rising, creating a cocktail of irony, ridicule and humiliation that is attractive to young people, particularly,” said Selma Muhić Dizdarević and Jitka Votavová,<sup>68</sup> the authors of the study on antigypsyism within the international project Remember and ACT! (Re-ACT), which concentrates on researching “old” concepts of hatred in their modern forms.

61 *Ibid.*

62 Ismael Cortés, ‘Hate Speech, Symbolic Violence, and Racial Discrimination. Antigypsyism: What Responses for the Next Decade?’, *Social Sciences* 10, no. 10, 2021, p. 360. Available at: <https://doi.org/10.3390/socsci10100360>.

63 *Ibid.*

64 Selma Muhić Dizdarević, *Comparative Report on the phenomena of online antigypsyism*, Re-ACT project/INACH, 2020. Available at: <https://react.inach.net/wp-content/uploads/2020/10/Re-Act-Comparative-report-on-the-phenomena-of-online-antigypsyism.pdf> (accessed November 2021).

65 *Ibid.*

66 European Commission, *Countering illegal hate speech online: 6th evaluation of the Code of Conduct*. Prepared by Didier Reynders, Commissioner for Justice. Factsheet, 7 October 2021.

67 OpCode. *Monitoring and Reporting Illegal Hate Speech: Shadow Monitoring Report – first edition*, 2020.

68 Project Re-ACT Press Release, *Antigypsyism and antisemitism online - resources, research, stakeholders, educational hub*, 10 August 2021. Available at: <https://react.inach.net/wp-content/uploads/2021/08/2021-08-10-Re-ACT-press-release-english.pdf> (accessed November 2021).

In a research project on online antigypsyism in Austria, the Czech Republic, Germany, France, Italy, Latvia, and Slovenia, three clusters in which the most common narratives related to anti-Roma hate speech online emerged: criminalisation, welfare chauvinism, and dehumanisation.<sup>69</sup>

It must be considered that besides the common negative attitudes towards Roma that we can find across Europe, there are specificities in different countries. One of the goals of this pilot research has been to discover typical anti-Romani online narratives in four countries: Albania, Serbia, Turkey, and Ukraine.

<sup>69</sup> Maren Hamelmann (Ed.). *Antigypsyism on the Internet*, The sCAN project, 2018. Available at: <http://scan-project.eu/wp-content/uploads/scan-antigypsyism.pdf> (accessed November, 2021).

## Context of hate speech and antigypsyism in the target countries

### Albania

In the sixth report on Albania by the European Commission against Racism and Intolerance (ECRI), adopted on 7 April 2020, it is stressed that “...*hate speech, especially against members of the Roma and LGBTI communities, is still far too often considered to be an acceptable feature of public debates.*”<sup>70</sup> One of the recommendations in the report was that authorities should publicly condemn incidents of hate speech, especially against those two most targeted groups.

There are two equality bodies in Albania relevant to the work of ECRI: the People’s Advocate (Ombudsman) and the Commissioner for the Protection from Discrimination (CPD).

Racist humour targeting Roma is a common feature in entertainment in Albania. At least two popular comedy TV shows broadcast on national television use mocking of Romani people and racist depictions of Romani characters based on stereotypes as entertainment. *Portokalli*, broadcast on Top-Channel, and *Al Pazur*, on Vision Plus, depicted Roma as people not to be trusted; they are uncivilized, uneducated, and always have a lot of children. The ERRC’s human rights monitor for Albania, Xhenson Cela, noted that in both of these television programmes “*the Romani characters are ridiculed for speaking in Albanian with a weird accent while the audience explodes with laughter. The comedy shows, directly or indirectly, put the Romani characters in an inferior position.*”<sup>71</sup>

The ERRC volunteers from Albania, as part of this project, submitted a complaint to Commissioner for Protection from Discrimination related to both shows on December 2020. The office of the Commissioner had not responded at the time of writing.

70 ECRI, *Sixth report on Albania*. Adopted on 7 April 2020. Available at: <https://rm.coe.int/report-on-albania-6th-monitoring-cycle-/16809e8241> (accessed November 2021).

71 ERRC, *Blackface, Stereotypes, and Prejudice: Albania’s Racist Comedy Shows*, 2021. Available at: <http://www.errc.org/news/blackface-stereotypes-and-prejudice-albanias-racist-comedy-shows>.

## Serbia

In the latest ECRI report on Serbia from 2017 it is stated that:

*Hate speech is increasingly disseminated via the Internet; football hooligans and their organisations also contribute to spreading hatred. The system of (self) regulation of the media is not working properly: the Press Council is too weak and social media operators do not prevent and remove hate speech. The application of the legislation against hate speech and violent hate crime is inefficient and there is no decisive action against the activities of racist, homo- and transphobic hooligan groups.<sup>72</sup>*

In the same report, the cause for the proliferation of inflammatory language and hate speech in Serbia is suggested to be because: “many media outlets are struggling to survive commercially following their recent privatisation, resulting in a growing “tabloidization” of the print media and an increase in the number of reality shows on television.”<sup>73</sup>

Similar conclusions are found in a recent Council of Europe study on the use of hate speech in Serbian media: “The existence of hate speech and discriminatory speech is very high in Serbia, especially against LGBTI persons, Roma, women and migrants, and the greatest responsibility for this situation have public institutions, politicians and media itself.” Interviews were conducted with the representatives of relevant bodies combating hate speech during November and December 2020, such as the Commissioner for the protection of equality, judges specialised in anti-discrimination law, the Ombudsman, Press Council etc.<sup>74</sup> It was also concluded that Serbia has a solid legal framework for the protection against hate speech, however accountability for its implementation is lacking.

In the same study it was pointed out that the Regulatory Body for Electronic Media (REM) does not act in accordance with its jurisdiction. In 2019, the REM rejected all submitted complaints as incomplete. Out of 167 applications, 162 referred to the content of the programs broadcasted by reality shows, which are recognised as hate speech.<sup>75</sup>

In this regard, it is also worth mentioning findings from the same source:

*The largest number of citizens’ applications refers to two televisions with a national frequency which broadcast reality shows (TV Pink and TV Happy). During 2020, in January alone, a total of 78 charges were filed against TV Pink in connection with reality programs, which are related to animal protection and hate speech. Until September 2020, REM did not file any criminal or misdemeanour charges against media service providers for discriminatory speech, violations of the protection of minors, and hate speech in electronic media, although this falls within its competence and with evident daily broadcasting of such content.<sup>76</sup>*

72 ECRI, *Fifth Report on Serbia*, CRI(2017)21. Adopted 22 March 2017. Available at: <https://rm.coe.int/third-report-on-serbia/16808b5bf4>.

73 *Ibid.*

74 Ivana Krstić, *Report on the Use of Hate Speech in Serbian Media*, Council of Europe, 2020.

75 Vida Petrovic Skero and Natasa Jovanovic, *Analysis of the Effect of the Work of REM, 2017-2020*, Slavko Curuvija Foundation, Belgrade, p. 25. (Taken from Ivana Krstić, *Report on the Use of Hate Speech in Serbian Media*, Council of Europe, 2020).

76 Vida Petrovic Skero and Natasa Jovanovic, *Op.cit.*, p. 53, 68. (Taken from Ivana Krstić, *Report on the Use of Hate Speech in Serbian Media*, Council of Europe, 2020).

## Turkey

Among the four targeted countries covered in this report, the situation in Turkey is the worst in terms of a comprehensive anti-discriminatory legislation, including legislation on hate crimes.

In the latest ECRI report on Turkey from 2016, it was noted that:

*Turkey has not ratified Protocol No. 12 to the European Convention on Human Rights and the grounds of ethnic origin, colour, language, citizenship, sexual orientation and gender identity are missing from several criminal-law provisions. The definition of hate crime is excessively narrow and the Criminal Code does not explicitly provide that racist and homo/transphobic motivation constitutes an aggravating circumstance. Some core elements of the anti-discrimination law are not in line with ECRI's recommendations and it does not provide for the necessary independence of the new Human Rights and Equality Authority, which is however vital. Concerns also persist with regard to the independence of the Ombudsman Institution.<sup>77</sup>*

In the same report, it was concluded that “Hate speech is on the rise and its increasing use by officials, including senior representatives of the state, is of major concern.” Even though Turkey does not collect data on racist and homo/transphobic violence, civil society reports point to a high number of such hate crimes including deaths of LGBTQI+ people, transgender persons, and members of other minority groups. Several mob attacks against Roma and Kurds have also been recorded.<sup>78</sup>

In a more recent study from 2020, high rates of hate crimes in Turkey have been related to the notion of ‘identity’ and its relevance to the state’s policies. Identities that had been designated through religious references in the Ottoman Empire were replaced in 1923 when the Turkish Republic was proclaimed, by a homogenous state identity of “Turkishness”. This categorisation found its expression in constitutions while citizenship was defined with reference to Turkish identity.<sup>79</sup>

Although Article 10 of the Turkish Constitution<sup>80</sup> lists a basic definition of discrimination, it does not include such fundamental concepts as “ethnic background, gender identity, and sexual orientation, all kinds of faith or lack thereof.” Hate speech in Turkey targets these groups intensively.<sup>81</sup>

77 ECRI, *Fifth report on Turkey*. Adopted on 29 June 2016. Available at: <https://rm.coe.int/fifth-report-on-turkey/16808b5c81> (accessed November 2021).

78 *Ibid.*

79 Human Rights Association, *Special Report on Hate Crimes and Recent Racist Attacks in Turkey*, September 2020. Available at: [https://ihd.org.tr/en/wp-content/uploads/2020/09/sr20200922\\_Hate-Crimes-and-Racist-Attacks-Report\\_Sept-2020.pdf](https://ihd.org.tr/en/wp-content/uploads/2020/09/sr20200922_Hate-Crimes-and-Racist-Attacks-Report_Sept-2020.pdf) (accessed November 2021).

80 Article 10: “Everyone is equal before the law without distinction as to language, race, color, sex, political opinion, philosophical belief, religion and sect, or any such grounds.” Available at: <https://www.anayasa.gov.tr/en/legislation/turkish-> (Taken from Human Rights Association, *Special Report on Hate Crimes and Recent Racist Attacks in Turkey*, September 2020).

81 Human Rights Association, *Special Report on Hate Crimes and Recent Racist Attacks in Turkey*, September 2020. Available at: [https://ihd.org.tr/en/wp-content/uploads/2020/09/sr20200922\\_Hate-Crimes-and-Racist-Attacks-Report\\_Sept-2020.pdf](https://ihd.org.tr/en/wp-content/uploads/2020/09/sr20200922_Hate-Crimes-and-Racist-Attacks-Report_Sept-2020.pdf) (accessed November 2021).

## Ukraine

In pre-war Ukraine, far-right groups and their activities have long posed a threat to the safety of minority groups.

In the last ECRI report on Ukraine from 2017, Roma are presented as the most frequent targets of racist violence by non-state and state actors. The report states: “...*racist violence committed by police continues to be reported as well as failure by police to intervene to stop racist or homophobic attacks.*” ECRI cite two mob attacks from 2014 and a pogrom from 2016 in the village of Loshchynivka in which more than 300 people took part in the violence.<sup>82</sup>

*“Although the head of the State Security Service stated in the monitored period that there are no radical right organisations registered in Ukraine, ECRI warns that there continue to be extremist organisations which manifest intolerance towards vulnerable groups and incite racial hatred.”<sup>83</sup> ECRI has also been informed that some of these groups, or individuals within them, have become involved in military action in the East of the country, thus gaining popularity for their openly ultra-nationalist agenda.”<sup>84</sup>*

The ERRC has monitored pogroms of Romani communities in Ukraine from 2011 – 2019, with many committed by far-right paramilitaries and state sponsored security agents. There were at least seven attacks on Romani communities in 2018 and 2019, in Lviv and Beregovo, with two Romani people murdered during this time. Several legal cases were taken by the ERRC and the NGO Chirikli, including a complaint against the National Police of Ukraine for discrimination and negligence.<sup>85</sup>

Offline violence during these pogroms went hand-in-hand with a strong social media presence from far-right groups, notably the neo-Nazi group formerly known as C14 (now known as ‘Foundation for the Future’, whose previous name was a reference to a known white supremacist slogan).<sup>86</sup> Many of the attacks were carried out in a staged manner and usually live streamed on social media. Beforehand there were warnings issued on Facebook, and ultimatums to leave the territory or face the consequences, furthering the echo chamber of hate speech against Romani people on Ukrainian social media. The attacks themselves were militarised and theatrical, with face masks and post-production music added to footage of the incidents. Messages were then sent to followers after the attack was finished declaring the area ‘cleansed’ of Roma.

The situation which existed in Ukraine prior to the war is an illustration of the influence of online media as a radicalisation tool for far-right movements, as well as demonstrating the need to hold online media accountable in order to prevent similar escalations.

<sup>82</sup> ECRI, *Fifth report on Ukraine*. Adopted on June 2017. Available at: <https://rm.coe.int/fifth-report-on-ukraine/16808b5ca8> (accessed November 2021).

<sup>83</sup> UN CERD, 2016. Taken from the ECRI fifth report on Ukraine.

<sup>84</sup> ECRI, *Fifth report on Ukraine*. Adopted on June 2017. Available at: <https://rm.coe.int/fifth-report-on-ukraine/16808b5ca8> (accessed November 2021).

<sup>85</sup> ERRC, *Written Comments of the ERRC concerning Ukraine For Consideration by the Human Rights Committee at its 129<sup>th</sup> Session (29<sup>th</sup> June – 24<sup>th</sup> July 2020)*. Available at: [http://www.errc.org/uploads/upload\\_en/file/5242\\_file1\\_ukraine-un-hrc-28-may-2020.pdf](http://www.errc.org/uploads/upload_en/file/5242_file1_ukraine-un-hrc-28-may-2020.pdf).

<sup>86</sup> “14 Words” is a reference to the most popular white supremacist slogan in the world: “We must secure the existence of our people and a future for white children.” The slogan was coined by David Lane, a member of the white supremacist terrorist group known as The Order (Lane died in prison in 2007). See “14 Words”, Anti-Defamation League. Available at: <https://www.adl.org/education/references/hate-symbols/14-words>.

# Methodology

## Data Collection

Trained volunteers from each participating country (Albania, Serbia, Turkey, and Ukraine) searched for online examples of hate speech targeting Roma (according to the agreed working definition for this project) in their local languages in the period from November 2020 until July/August 2021. They searched online news portals and social networks such as Facebook, Twitter, YouTube, Instagram, TikTok, and Telegram.

The most searched platforms in each country differ due to varying user habits, as well as the fact that each volunteer searched on the networks they were personally most active on. Discriminatory content could be in text format or visual presentations such as pictures, memes, and videos, and could be shared online through a profile, page, post, or a comment. The volunteers searched known “hot spots” for online hate speech, as well as using keywords to uncover results.

More about the search methodology is presented in the Results section for each country.

The volunteers’ main task was to report examples of hate speech using available reporting tools on social media platforms, make a print screen/screenshot of the reported content, and to store it in the corresponding country folder.

Volunteers were advised to keep a record of the source and the status of the reported content. The source refers to where on the social network/online platform the example was found, and the status refers to whether a given social network/online platform removed the reported content or not.

As the results section of this report will show, the status of the reported content was not always easy to follow. Except for Facebook, social networks rarely, if ever, notify users about the status of the reported content. Additionally, some online media platforms do not have an option for reporting hate speech at all. This resulted in the information about the status of the reported content being incomplete in some country results.

## Processing data

A minimum of 300 reported cases of hate speech were collected by each national group. Each sample was analysed according to the following characteristics: status (successful, unsuccessful, unsorted); source (Facebook, YouTube etc.), type of content (comment, post etc), and main category related to the project’s hate speech definition.

National coordinators assigned codes to each collected hate speech case (usually a screenshot) and a Python algorithm was developed and applied to process the data.



Analysis of the content was undertaken according to the project's working definition of hate speech, which divided recorded hate speech into 4 categories of content:

Content which uses dehumanising or inferior language.

Content encouraging or glorifying violence.

Content which uses taboo language/ethnic slurs.

Content which incites hatred or discrimination.

Since in many cases reported content contains several of these categories at once, volunteers were instructed to give priority to the most damaging category present. The order for this is as follows:

1. Encouraging or glorifying violence;
2. Dehumanising or inferior language;
3. Inciting hatred or discrimination;
4. Using taboo language.

In the Results section of this study, this is referred to as the 'main category of the project's hate speech definition'.

While coding it proved possible to include other categories of the hate speech definition when present. A formula for prioritising the codes of the main category of the hate speech definition was included in the algorithm applied during the data processing in order to rectify possible mistakes that had been made during coding.

## Qualitative data

As well as quantitative data (frequency of the codes assigned to the collected samples of the anti-Roma speech), qualitative research methods were also applied. Qualitative data was collected through the structured questionnaires before quantitative analysis was performed, and then semi-structured interviews with the national volunteer coordinators which were held after the quantitative data was processed. Additionally, since it was assumed that online content would contain many stereotypes and prejudices about Romani people, national coordinators were asked to estimate which categories of prejudices and stereotypes would be the most present amongst their data (by percentage).

A framework of categories related to typical stereotypes and prejudices towards Roma was created using existing research on the common stereotypes and prejudices targeting Roma, the ERRC's extensive experience in combating anti-Roma discrimination, and insights provided by the volunteers regarding this topic in their respective countries. This framework served as

a basis for national coordinators to estimate which categories of stereotypes and prejudices toward Roma are the most present in their collected data. The categories were divided into types of stereotypical content which:

- relates to criminal and violent behaviour (e.g., Roma steal children, they are thieves, they are a threat to public safety).
- relates to work (e.g., Roma don't want to work; they only want to beg).
- relates to personal characteristics (e.g., Roma are dirty, ugly, stupid, cannot be trusted etc).
- relates to way of life or culture (e.g., Roma are inadaptable, unchangeable, respect only their own rules).
- relates to law/the state/ the state's policies (e.g., Roma are protected by the law, the majority population is discriminated against and Roma are privileged, Roma use all the social aid budget so there is nothing left).
- relates to sexual behaviour (e.g., Roma are promiscuous, they have too many children).
- relates to religion or morality (e.g., Roma are blasphemous, Roma are apostates, Roma are amoral).

The Turkish team conducted an analysis of the typical narratives by counting key words and themes present in each case, and coordinators from Albania and Serbia made estimations based on the suggested categories. An analysis of this kind was not conducted in Ukraine.

## Limitations of the research

The obvious limitation of this research is also the main element which makes it most empowering as a form of activism; that volunteers were engaged to carry out the project, rather than experienced researchers.

One of the main limitations is that data was not systematically collected. Due to the nature of the project, which relied on volunteered time, the data could not be uniformly collected from all countries and so a comparative analysis is not possible. Rather, each data set, and qualitative conclusions, portray a small but representative snapshot of online hate speech against Roma in each country.

There was also a disproportion in the distribution of different types of reported content (comments, posts, profiles, videos, images, pages) with comments dominating in all four samples.

There were also omissions in the classification of collected data related to the status of the reports. Due to large amounts of reported content being unclassified (aka not known whether it was removed or not) a complete statistical analysis was not possible for all the countries, with the exception of the data collected in Serbia, and partially from data collected in Turkey. Therefore, the quantitative analysis of reports in the other countries may be viewed as evidence for that body of reports only, and not evidence of overall trends for social media platforms' moderation successes and failures.

Since this was the first research project monitoring online hate speech targeting Roma in relation to the social media platforms' moderation in targeted countries, lessons learned in the process will serve to improve research procedures in the volunteer teams' next monitoring exercises.

## Findings

### Albania

The Albanian team were engaged in the process of finding and reporting online hate speech content targeting Roma from November 2020 until June 2021.

Key words used for searching anti-Romani narrative were: #gabel #jevg #magjup #zezak #gabelet #jevgjit #arixhi #harixhi #arixhinjt.

All the above words are used as ethnic slurs for Romani people. The words *jevg*, *jevgjit*, and *zezak* are slurs in Albanian referring to the Balkan Egyptian community.<sup>87</sup>

The Albanian team also identified and monitored a number of groups (Qesh pa kufi, odaj, odaj) and pages (Legjendat e Humorit, Thjesht Humor lal, Humor.al) with a large number of followers on Facebook and Instagram where anti-Romani content is often expressed through inappropriate humour and memes aimed at mocking Roma.

The majority of the 408 reported cases in the Albanian sample came from TikTok (274), followed by examples from YouTube (55) and Facebook (54). There were also 25 examples from Instagram.

Initially, volunteers in the Albanian team were divided into smaller groups in order to cover all social networks and platforms. However, they decided to narrow down their monitoring to just the above 4 networks since those were where they had been finding the majority of the examples of hate speech targeting Roma.

The coordinator of the Albanian volunteer group, Roxhers Lufta, reported that in the initial monitoring period volunteers did not find examples of hate speech on Twitter. This led to them stopping monitoring of that network in order to focus on others where hate speech was more prominent. They also noticed that it was difficult to find examples of hate speech on online news portals. Lufta believes this is because news portals take care not to use discriminatory language against Roma.<sup>88</sup>

When asked about the comparatively high number of cases from TikTok and whether anti-Roma speech is more prevalent there or if it's due to other reasons, Lufta suggested that the main reason for this result is the high popularity of TikTok in Albania.<sup>89</sup> TikTok also tends to

<sup>87</sup> Balkan Egyptians are a distinct ethnic group present in the Southern Balkans, most prevalently in Kosovo, North Macedonia, Albania, Serbia, and Montenegro. They are culturally distinct from Roma, although they are often erroneously grouped together by outsiders. More information is available at: [https://www.coe.int/t/dg4/education/ibp/source/PS\\_1\\_10.5.pdf](https://www.coe.int/t/dg4/education/ibp/source/PS_1_10.5.pdf).

<sup>88</sup> Interview with Roxhers Lufta, October 2021.

<sup>89</sup> Interview held in October 2021.

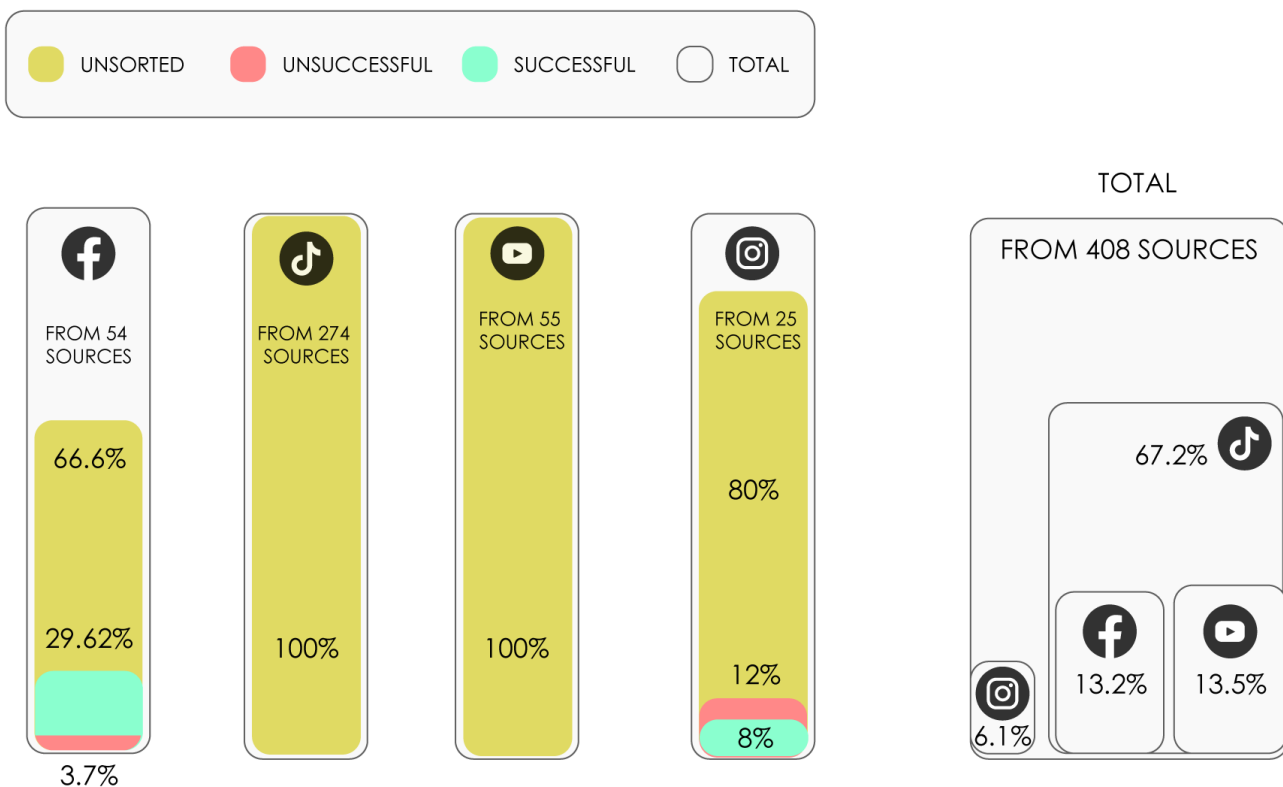
## FINDINGS

appeal to a younger demographic, which the Albanian volunteers fall under, and so the higher number of cases on TikTok may correspond to the volunteers' activity levels on that social media network.

The vast majority of the reported content was in the form of comments (347). There were also 58 reported posts, 2 pages, and 1 profile.

### RELATION BETWEEN THE SOURCE AND REPORT STATUS

In the Albanian sample, the vast majority of reports are classified as unsorted (385) as in many instances volunteers were not successful in keeping track of whether the offending content was removed from the platform or not. It is only known that the reports were successful (removed) in 18 instances, and unsuccessful (not removed) in 5 instances.



All reports from TikTok and YouTube are unsorted. The majority of the successful reports from the Albanian sample are from Facebook (16 out of 18 successful reports). There are also a small number of reports (5) from Instagram where the status of the reports is known. However, the majority of reports are also unsorted for these two social media networks.

Lufta explained that in the process of coding saved reports the volunteers checked all unsorted reports from Facebook and Instagram, and the comments or posts in question were

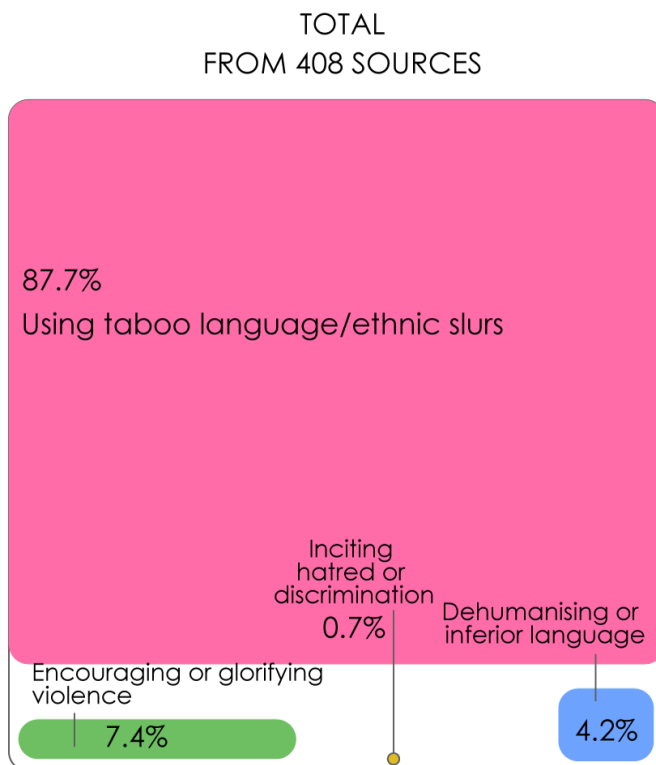
no longer visible. However, since they did not receive notifications regarding the status of these reports, they decided to code them as Unsorted.<sup>90</sup>

This indicates that a more accurate result would demonstrate a higher number of successful reports on Facebook and Instagram. In any case, conclusions about the efficiency of social media networks in removing reported hate speech content cannot be drawn due to the incomplete information.

Lufta stated that among the 4 sources for their sample, Facebook was the most responsive in sending notifications regarding the status of reported content but that notifications were not sent every time. He suggests that Facebook should work on recognising all variations of ethnic slurs used in Albania for Romani people. The worst social media network for sending feedback was TikTok, based on the experiences of Albanian team. “*Not a single notification was sent to us for all the reports we have made*”, said Lufta.<sup>91</sup>

### TYPES OF HATE SPEECH

When perceived through the main categories of the project’s hate speech definition, cases ‘Using taboo language/ethnic slurs’ dominate the sample with 87.75%:



<sup>90</sup> *Ibid.*

<sup>91</sup> *Ibid.*

## FINDINGS

The main type of hate speech present in the Albanian sample is usage of ethnic slurs, and this is the case across all four of the main sources of reports. ‘Encouraging or glorifying violence’ was present in 30 collected reports (19 from TikTok and nine from Facebook). ‘Inciting hatred or discrimination’ was the least present category.

### FACEBOOK - 54



### INSTAGRAM - 25



### TIKTOK - 274



### YOUTUBE - 55



Lufta is not surprised that usage of ethnic slurs constitutes such a large percentage of the online content, as he feels using ethnic slurs to refer to Romani people is “normal” in Albania and Kosovo. He believes that the majority of non-Romani people are not aware that these names (*Gabel* and its variations) are offensive towards the Romani community and that not much is known about Romani people in general. He also stresses that many Roma use these ethnic slurs because even they are not always aware of the pejorative meaning.<sup>92</sup>

## TYPICAL ANTI-ROMANI NARRATIVES

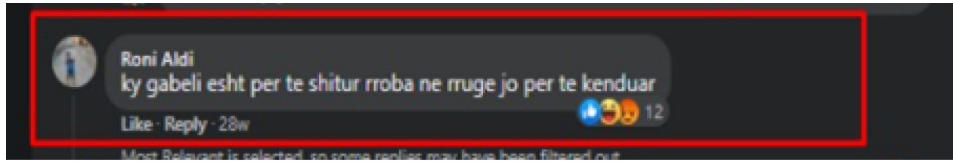
In addition to identifying examples of the project’s main definition of hate speech, the Albanian team were also asked to examine online content which does not fall under this main definition of hate speech but which is still racist and prejudiced.

The team discovered that the dominant example of prejudice appearing in their sample related to work (around 40%) with statements such as “Roma don’t want to work”; and “only begging”. In around 25% of reports examples of prejudices related to personal characteristics were present, and in around 10% of reports examples of prejudices related to sexual behaviour were present. The remaining 25% consisted of a mix of prejudicial comments.<sup>93</sup>

<sup>92</sup> Interview with Roxhers Lufta, October 2021.

<sup>93</sup> Questionnaire, August 2021.

This is a comment on a video of famous Romani musician Mandi Nishtulla:



**Translation** – this ‘gabel’ is for selling clothes, not for singing

Calls to exterminate all Roma are also present in the Albanian sample:



## Translation of the marked comment:

Qazim Mulleti said “well extinguish this race”.

Qazim Mulleti was the Mayor of Tirana from 1942-1944 known for racist and violent policies toward Romani communities. Source: Facebook. Status: unsuccessful.

In Albania, making fun of Romani people as a form of entertainment is common. The Albanian team highlighted one example from TikTok that is a good illustration of this. A video was created where an adult records a conversation with a Romani boy. A translation of their dialogue is as follows:

**Adult :** *Aren't you afraid that Covid-19 will infect you?*

**Roma boy:** *I am not afraid because I am ‘Gabel’*

**Adult:** *And who can be affected by the virus?*

**Boy:** *“Magjup”*

## FINDINGS

The boy is saying that he cannot be affected by Covid-19 because he is ‘*Gabel*’, not ‘*Magjup*’. The joke here is that both ‘*Gabel*’ and ‘*Magjup*’ are ethnic slurs referring to Romani people: ‘*Gabel*’ is used in Albania and ‘*Magjup*’ in Kosovo. The boy is not aware of this, revealing his lack of education.

The Albanian team found that 720 people used the audio of this conversation to make a TikTok video lip syncing along.<sup>94</sup> It was mostly young adults participating in this TikTok trend, however rather disturbingly there were also a number of videos made by children.

During 2020, the Albanian volunteers identified two popular comedy TV shows; Portokalli and Al Pazar, that depict Romani characters based entirely on stereotypes and prejudices for the sake of entertainment. On December 2020 they submitted an “Open Paper” to the Commissioner for Protection from Discrimination related to both shows. No response has been received at the time of publication.

To conclude, the main characteristics of anti-Romani narratives identified by the Albanian team are:

- large overrepresentation of ethnic slurs; and
- use of racist humour.

## CONCLUSIONS

The majority of reported content in the Albanian sample came from TikTok (274), followed by examples from YouTube (55) and Facebook (54). There were also 25 reports from Instagram. Twitter and online news portals were also examined and found not to contain significant numbers of hate speech targeting Roma.

Conclusions about the efficiency of networks in removing reported hate speech content could not be drawn due to the incomplete information related to status of the reports. A likely contribution to this issue is the fact that TikTok, as the most represented social media network, does not send notifications regarding the status of reported content.

The ethnic slur ‘*Gabel*’ and other slurs referring to Romani communities in Albania dominate the sample with 87.75%. This type of anti-Romani narrative is the most common from all four main sources of reports.

In the Albanian sample, the most represented prejudices are related to work (around 40%) followed by prejudices related to personal characteristics (around 25%), and prejudices related to sexual behaviour (around 10%). Prejudices and stereotypes towards Roma are found to be particularly reinforced through widespread and accepted racist humour.

<sup>94</sup> This can be viewed at: <https://vm.tiktok.com/ZMdVST5yW/> (accessed September 2021).



## RECOMMENDATIONS

The Albanian government and public authorities should:

- Ensure full implementation of existing anti-discrimination laws and related treaty commitments. Raise awareness about antigypsyism, and challenge it through anti-racist training for political leaders, public officials, law enforcement and criminal justice professionals with a view to mainstreaming civic values based on inclusion and the protection of fundamental rights for all.
- Refrain from, and condemn all instances of anti-Roma hate speech and the usage of disparaging ethnic slurs.
- Cooperate with NGOs, equality bodies, media outlets and celebrities and publicly campaign to counter anti-Roma narratives and hate speech in broadcast, print and online media.

Social media networks should:

- Revise their community standards to ensure that they take account of the specificities of antigypsyism, and to ensure that community standards are fully applied to sanction online hate speech.
- Upgrade machine learning systems to be more adept at recognising Albanian-language ethnic slurs which refer to Roma: #gabel #jevg #magjup #zezak #gabelet #jevgjit #arixhi #harixhi #arixhinjt.
- YouTube should engage with civil society to increase familiarity among activists with the YouTube Trusted Flagger program, and explore how best to collaborate in taking down hate content.
- TikTok needs to improve its response rates in reviewing flagged material and removing hate speech content. Having signed up to the EU Code of Conduct on online hate speech, TikTok should ensure these commitments apply to Albania, and to Roma in Albania

## Serbia

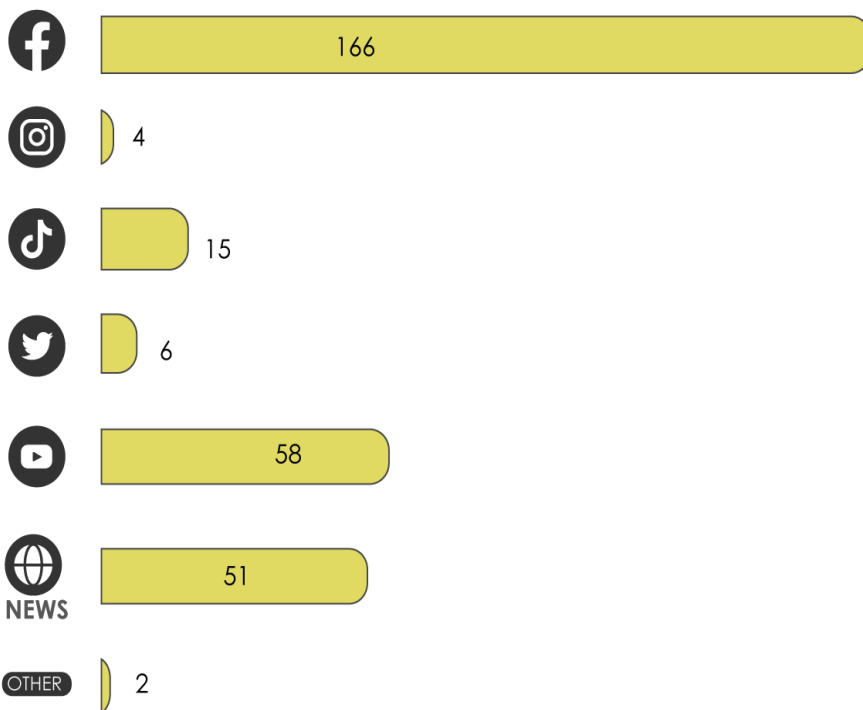
The Serbian team searched for and reported online hate speech content targeting Roma from November 2020 until August 2021. The majority of reports were flagged and data collected in the period between June and August 2021.

Key words used for searching for anti-Romani narratives on the four targeted social media networks were: #cigan; #ciganin; #ciganski; #ciganska; #ciganstura; #ciganstina; #mandov; #kalafonac; #kalafonci; #kalafonka; all of which are ethnic slurs against Romani people in Serbia. The volunteers also searched for #romski; #romska; #romsko as key words in order to bring up as many relevant results as possible.

The volunteers also examined other sources for anti-Romani narratives, including Facebook pages with a nationalist and right-wing orientation, online news portals' articles and posts concerning Roma on related Facebook pages, and YouTube channels related to reality TV shows.

## FINDINGS

The Serbian team collected 302 examples of hate speech, with the majority found on Facebook (166). Other highly represented sources were: YouTube (58 cases) and online news portals (51 cases). All the collected data represents reported examples of hate speech, except in some cases from online news portals where there was no reporting function.



Aleksandar Smailović, the volunteer coordinator for Serbia, explained that one of the reasons why Facebook represents the highest number of reports is because it is the social media network most used among the volunteers who participated in the research.<sup>95</sup>

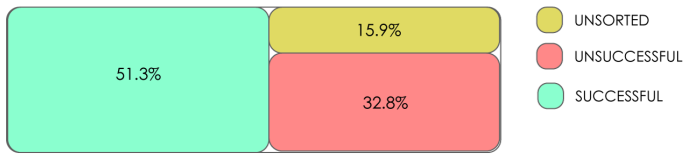
In the Serbian sample, the vast majority of reported content was in the form of comments (252). There were also 45 posts, 3 reported pages, and 2 personal profiles. The majority of the reported content was in text format (92.7%) with only 7.3% in the form of a picture or video.<sup>96</sup>

As can be seen in the table below, out of 302 total cases 155 were successful, 99 unsuccessful, and 48 unsorted. This is the only country result out of the four targeted countries where there are more successful reports than unsuccessful. The proportion of unsorted data is also within acceptable limits.

<sup>95</sup> Interview with Aleksandar Smailović, September 2021.

<sup>96</sup> *Ibid.*

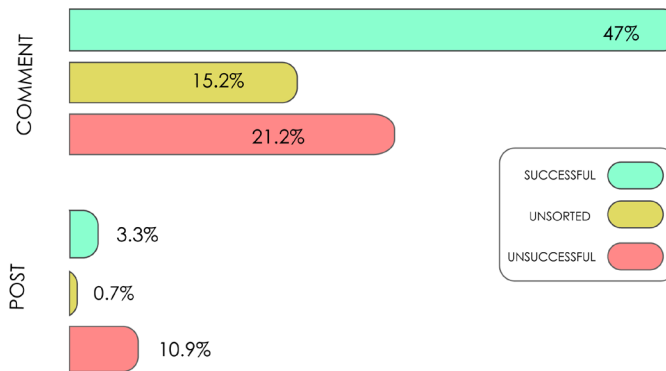
TOTAL - 302



The data from the Serbian team is the most accurately classified, as more efficient classification techniques were used and the national coordinator in charge of processing data was more experienced in the research techniques.

**RELATION BETWEEN STATUS AND THE TYPE OF CONTENT**

This analysis was only performed for the Serbian sample as the accuracy level of the data classification was high enough to provide useful insights. The graph below shows the proportion of successfully removed reported comments is almost double that of the unsuccessfully reported comments. However, the proportion of unsuccessful reported posts is much higher than successfully reported posts. This result indicates that social media networks are more reluctant to remove posts than comments under posts.



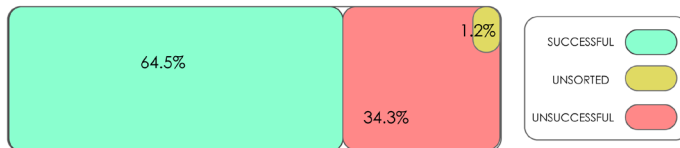
**RELATION BETWEEN THE SOURCE AND STATUS**

The majority of data in the Serbian sample came from Facebook (166). There were double the number of successful reports than unsuccessful reports from this social media network. Only 2 reports were unsorted.

The table shows results regarding the status of reports filtered by Facebook only:

## FINDINGS

FACEBOOK  
TOTAL - 166



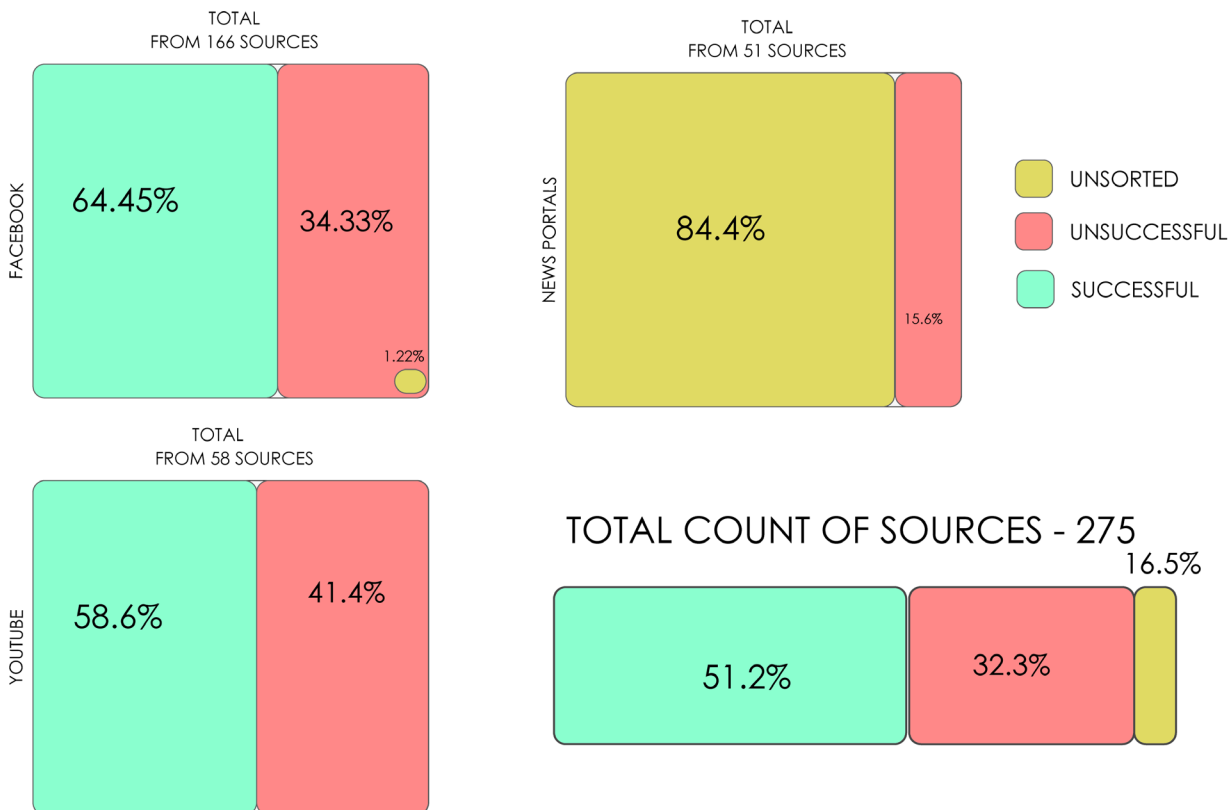
This result is encouraging since it shows that Facebook does process reports of content which violates community standards and does implement sanctions; by removing content recognised as hate speech.

Through analysing the reports from the three most represented sources in the Serbian sample (besides Facebook, these are YouTube and online news portals) an interesting result was uncovered in relation to online news portals: this source produced the majority of the unsorted cases and not a single successful report.

Out of all the unsorted cases from the three main sources, 95.56% of the unsorted reports stemmed from reported content found on online news portals. This result confirms observations by the Serbian volunteers that in many cases the comment section of the online news portals offers no reporting function. If there is the option to report a comment, no action seems to be taken.

The results from YouTube show that there is no significant difference between the proportion of successful and unsuccessful reports on this social media network.

The graph below shows the status of the reports from the three most represented sources of collected data:



## EXPERIENCE WITH ONLINE SOURCES AND TOOLS FOR REPORTING

Aleksandar Smailović, one of the volunteer coordinators for Serbia, explained that Facebook usually sends notifications to the reporter regarding the status of the reported content. This is in contrast to YouTube, which Smailović said *“did not send any notifications at all. So I kept the links where I found the content in question so that I could come back in three to four days to ‘manually’ check if the content had been removed or not.”*<sup>97</sup>

The volunteers found examples of hate speech on the following Serbian online news portals: “Espresso”, “Telegraf”, “Kurir”, “Alo!”, and “Novosti”. The examples were discovered in comments in the comment section below articles related to Roma.

Dragana Kokora,<sup>98</sup> one of the volunteer coordinators for Serbia, shared her experiences applying community standards at these online news portals where examples of hate speech were found. She explained that Telegraf.rs offers an option for leaving comments, but not for reporting comments; it only has a minus symbol (to indicate you do not like the comment) and plus symbol (to indicate you like the comment) as a type of reaction. Alo.rs and Novosti.rs also do not have an option for reporting comments, only the minus and plus reaction symbols.

Kurir.rs offers an option for reporting comments but only for registered users (the user needs to create a profile to report a comment, the same as for leaving a comment), but this function does not work well. When clicking the button to report a comment, a pop-up screen appears which requires log in and, after entering the email address and password, it returns you to the comment section without any confirmation that the comment was reported.

Espresso.co.rs is the only online news portal that offers all options; leaving comments, reporting comments, plus and minus reactions, and responding to existing comments. All the options work properly.

The example below displays a reported comment on the Espresso.co.rs portal. In the comment, the user is referring to a *“German leader who came to power in 1933”* and his *“genius”* since he conducted *“the first environmental movement to cleanse the planet of trash, degenerates, and moneylenders”*. The comment was left in response to an article with a sensationalist title: *“God created us, Gypsies, for music”* (in Serbian: *Mi Cigani smo od Boga stvoreni za muziku*).



### Translation:

The more knowledge and experience I have gained, the more I understand the genius of the German leader who came to power in 1933. That was the first environmental movement - to clean this planet from trash, degenerates, and moneylenders.

97 *Ibid.*

98 Interview conducted on 5 October 2021.

## FINDINGS

It is obvious that the comment refers to Hitler and his ideology to exterminate Romani people. It was reported in December 2020, but it was not removed. It can be still found on the news portal.<sup>99</sup>

### EXPERIENCES WITH TIKTOK

Although there was not as much hate speech content identified on TikTok, volunteers from the Serbian group brought particular attention to the trends they were encountering on this social media network.

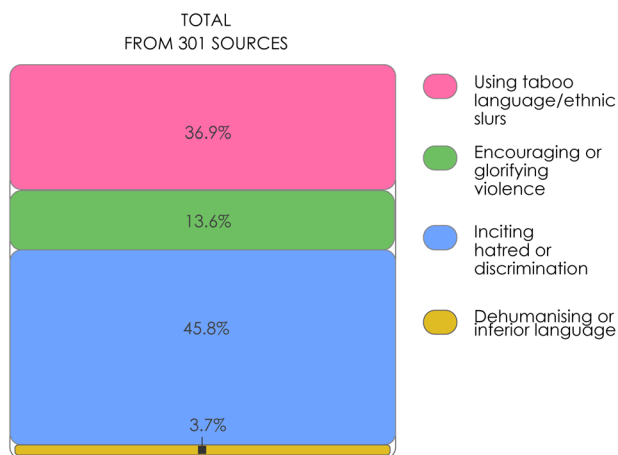
In particular, they noted content that degrades Romani people through humour and sarcasm. This content is in the form of short videos, recorded by the users of this social platform who use the lip sync technique to imitate already existing videos (mostly from the DNA TV show) whose subjects are Roma with poor knowledge of the Serbian language.

Moreover, the volunteers came across a number of vulgar jokes about Roma, but also insults directed towards users who are Roma or look Romani. What is encouraging in this regard is that other TikTok users can be seen to criticize these comments or posts; essentially applying community standards to their peers.

It was also observed that comments are removed from TikTok more quickly and frequently if recognised as hate speech, in comparison with reported videos and posts.

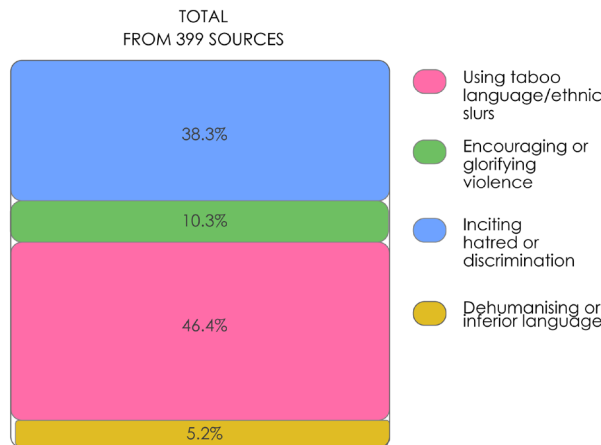
### ANALYSIS OF CATEGORIES OF HATE SPEECH

If the collected examples are analysed with regard to the project's main category of hate speech definition, it can be seen that examples inciting hatred or discrimination dominate the sample (45.8%). This is then followed by examples 'Using taboo language/ethnic slurs' (e.g. the word '*cigan*' in Serbian and its variations) with 36.9%. 'Encouraging or glorifying violence' represents 13.6% and 'Dehumanising or inferior language' represents 3.7%.



<sup>99</sup> Available at: <https://www.espreso.co.rs/vesti/drustvo/650687/nebojsa-saitovic-nesa-saita-crne-mambe-dzipsi-kings-saban-bajramovic-grad-nis-romi/komentari?fbclid=IwAR2d094NTwkCjVcwZrEwXEQAvqFjVmi82xMwsYGAZ2HuN4HSFtPg9Y1UljK>.

The majority of reported examples from Facebook fall into the ‘Using taboo language/ethnic slurs’ category of the project’s hate speech definition, while among cases from YouTube and online news portals ‘Inciting hatred or discrimination’ dominated. If all categories of the project’s hate speech definition are counted, regardless if these elements are perceived as main category or not, it can be seen that ‘Using taboo language/ethnic slurs’ dominates the sample:



It can be concluded that usage of the ethnic slur ‘*vigan*’ and its variations is overly present in online public spaces, despite its pejorative meaning and the fact that it is a politically incorrect way to refer to Romani people.

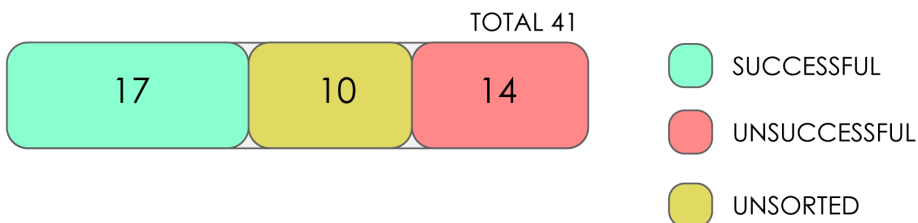
**HATE SPEECH DEFINITION AND STATUS**

Since the Serbian data is the most accurately classified as regards the status of the reports, a detailed analysis could be performed in order to uncover characteristics of digital content perceived by online networks/platforms to violate community standards.

**ENCOURAGING OR GLORIFYING VIOLENCE**

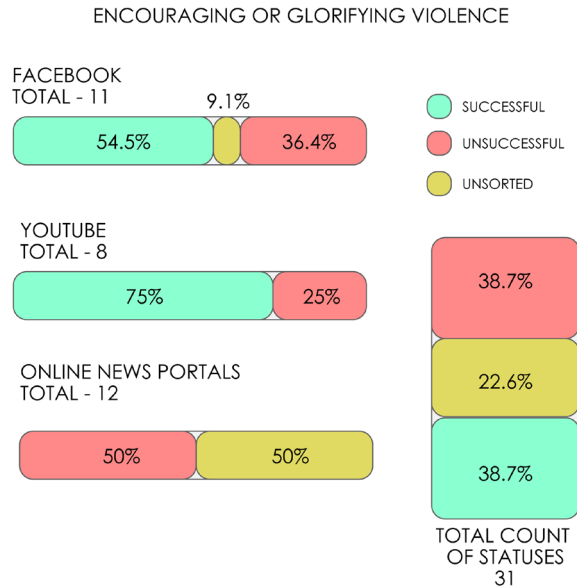
The table below shows the status of the cases coded as belonging to ‘Encouraging or glorifying violence’ as the main category of the project’s hate speech definition. There is not much variation between the number of successful (17), unsuccessful (14), and unsorted (10) cases.

**FREQUENCY**



## FINDINGS

There is not one dominant source of reports with this type of hate speech narrative. The table below shows examples of ‘Encouraging or glorifying violence’ in relation to the status of the reported content for three of the most represented sources in the Serbian sample; Facebook, YouTube, and online news portals.



There were also eight reported examples from TikTok labelled as ‘Encouraging or glorifying violence’. Of these five were successful, one unsorted, and two unsuccessful.

Despite the expectation that hate speech examples containing elements of ‘Encouraging or glorifying violence’ would be recognised as hate speech and so removed, in the majority of cases, this was not confirmed in the Serbian sample.

Further case analysis of unsuccessful reports on social media networks was performed to discover possible answers to the question of why some reported content is not removed in spite of expectations.

The example below is a post from TikTok that was not removed, despite the clear violent message:



### Translation:

FUCK YOU GYPSY MOTHER ALL OF THEM SHOULD BE BURNED AND KILLED SO THAT NEITHER THEM AND MIGRANTS DO NOT LITTER ANYMORE THEY TREAD THE SERBIAN FLAG IN THE MIDDLE OF SERBIA.



One of the possible reasons for the content not being taken down is that it is in the form of a post, and it has been shown that networks are more reluctant to remove posts.

The following are examples of unsuccessful reports with a narrative seen to be encouraging violence from Facebook:

### Translation:

In the quarry at 11 Bezirk, next to the Simmering cemetery, most of the people were Roma... the consequences of Adolf... 4 million Roma and 2 million cubic meters of gas.



In both examples there are several thoughts expressed in one grammatically incorrect sentence. In the first example, the user merged the words *Rom* (Roma) and *Cigan* (Gypsy) into a derivative “rom\*igan”. In the other example, Roma are mentioned in the context of the Holocaust with smile emoticons afterwards (both happy and sad emoticons, which adds to the ambiguous meaning of the message expressed in the text). It can be concluded that good knowledge of the language and the context is necessary in order to understand the true meaning of online content, especially when the content is expressed in the form of a symbol (picture), irony, sarcasm, euphemism and other figurative forms.

To conclude, digital content that clearly encourages violence against Roma were usually removed from Facebook, but examples with figurative or ambiguous meanings were not recognised as violating community standards.

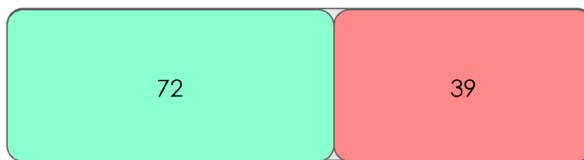
As previously discussed, the online news portals covered for this research did not have tools for reporting hate speech or, if they did, no action was taken against reported content. This data all contributes to the conclusion that the rate of removing content that promotes violence is not higher in the Serbian sample.

## USING TABOO LANGUAGE/ETHNIC SLURS

Digital content containing ethnic slurs (in Serbian this is particularly the word ‘*Cigan*’ and its derivatives, such as ‘*Ciganštura*’) was also reported. Below are the numbers of examples containing only this category of the project’s hate speech definition. There were 72 successful cases and 39 unsuccessful cases.



## FINDINGS

FREQUENCY  
TOTAL - 111



PERCENT TO WHOLE SAMPLE  
TOTAL - 36.9%

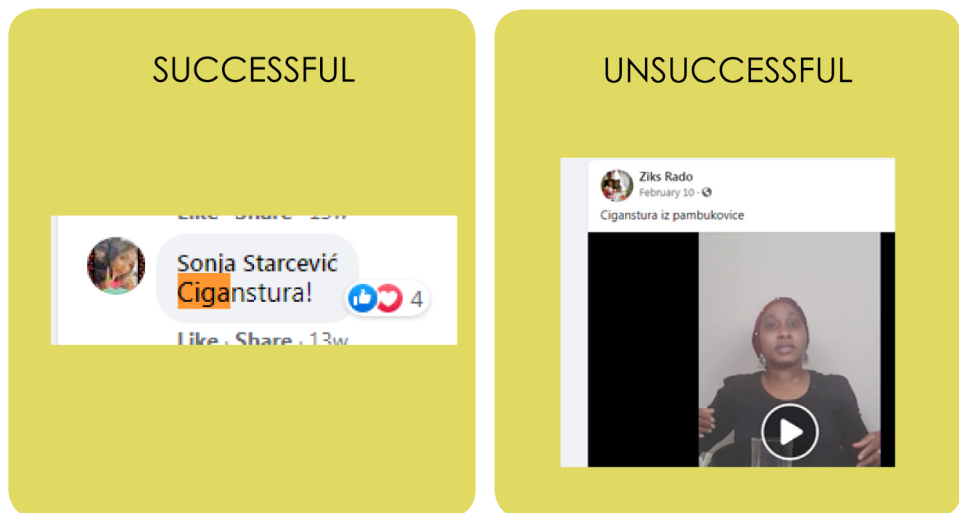


 SUCCESSFUL  
 UNSUCCESSFUL

In the Serbian sample, there were no unsorted cases with this type of narrative. By far the most numerous examples were from Facebook (94), and none were identified from online news portals.

From Facebook there are twice as many successful reports (63) than unsuccessful (31). Despite Facebook considering ethnic slurs to be a violation of community standards (and removing many such instances), in a significant number of cases Facebook moderators decided content using ethnic slurs did not violate community standards and the content was not removed.

To illustrate, both examples below contain the key word '*Ciganstura*'. The first case was removed while the second was not:



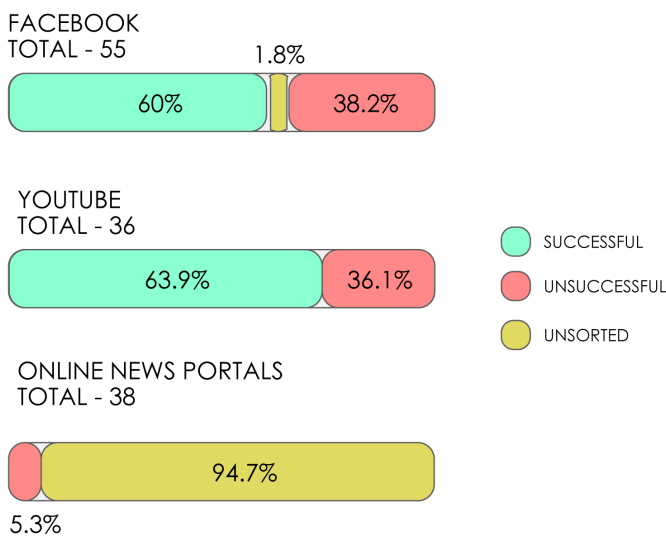
What differs between the two examples is that the first one was a comment while the second one was a post with video. It has already been observed that posts are less likely to be removed than comments.

## INCITING HATRED OR DISCRIMINATION

Examples where ‘Inciting hatred or discrimination’ is the most represented category of the project’s hate speech definition were found to be removed in 59 cases, not removed in 42 cases, and unsorted in 37 cases.

Examples of this type of anti-Romani narrative were present in all online sources. The table below shows examples of ‘Inciting hatred or discrimination’ in relation to the status of the reported content for three of the most represented sources in the Serbian sample; Facebook, YouTube, and online news portals:

### INCITING HATRED OR DISCRIMINATION



For Facebook and YouTube, examples of this type of hate speech have very similar numbers; in pprox.. 60% of cases the reported content is successfully removed but in pprox.. 40% it is not. Examples from online news portals commonly include examples of this type of hate speech, however the reported content is almost all unsorted for this research as those sources do not regularly apply community standards.

These numbers relating to the effectiveness of applying community standards by various social media networks confirm insights from Aleksandar Smailović, coordinator of the Serbian volunteer group:

*“The effectiveness of content removal is not consistent. Today, on the same (social) platform a comment with a stereotypical narrative (e.g. “Gypsies are dirty thieves”) will be removed, tomorrow “Gypsies, go back to India!” will remain.”<sup>100</sup>*

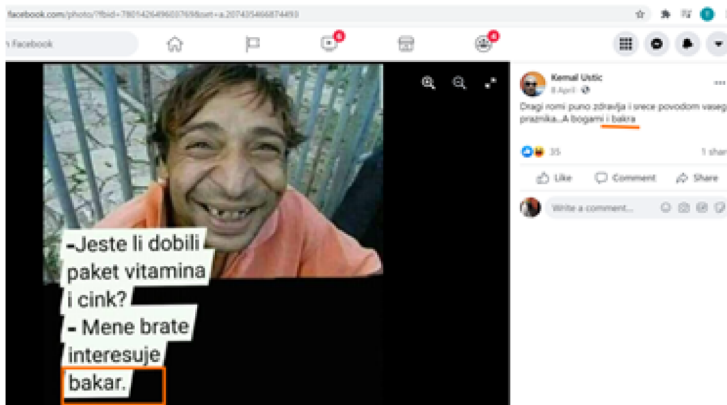
<sup>100</sup> Interview with Aleksandar Smailović, October 2021.

## TYPICAL ANTI-ROMANI NARRATIVES

Based on his extensive searches through online spaces, Aleksandar Smailović outlined the recurring themes in anti-Roma narratives in Serbia:

*“The general impression is that hate speech in the form of calling for violence and persecution of Roma is not as common as the speech full of stereotypes and negative prejudices in which Roma are being humiliated and perceived as primitive and uncultured compared to majority population. Additionally, we often saw comments claiming that Roma are protected even though they are ‘harmful’ to society; that they are privileged in relation to the majority population, in terms of education and social assistance, and that actually it is the majority which is discriminated against, not the Roma.”<sup>101</sup>*

The following example from Facebook displays this attitude:<sup>102</sup>



### Translation:

- Have you received a package of vitamins and Zinc?
- Dude, I am interested in copper.

Another observation is that channels dedicated to reality entertainment shows like “Zadruga” and “DNA” provide a ‘favourable climate’ for posting racist anti-Romani content. Volunteers investigated particularly how the reality show DNA functions as a virtual ‘nursery’ for the production of online racist content (short ‘funny’ videos, memes, comments), shared on Facebook, YouTube, and especially on TikTok which stereotypes and humiliates Roma.

As is shown in the statistical analysis related to main category of the hate speech definition in the Serbian sample, the majority of the reported content is categorised as that which incites hatred or discrimination (45.8%); and content which uses ethnic slurs such as ‘Cigan’ (36.9%). Another way of examining the main characteristics of the online anti-Romani speech represented in the Serbian sample was to analyse it through the lenses of prejudices and stereotypes. This is result of this analysis:

In terms of prejudices and stereotypes, the most prevalent narratives in the Serbian sample are:

- 22.2% - related to criminal and violent behaviour – typically referring Roma as thieves, or threats to public safety etc.

<sup>101</sup> Interview with Aleksandar Smailović, October 2021.

<sup>102</sup> This is an allusion to the packages of vitamins and zinc that the Government sent to all elderly people as a social measure during the Covid crisis.

- 20.9% - related to state policies and laws allegedly favouring Roma – statements such as: *Roma are protected by law; majority is discriminated and Roma are privileged; they are using all social help so there is no help for others.*
- 19.6% - related to sexual behaviour – statements such as: *they only know how to make kids; they f..k like cattle....*
- 15.7% - related to work and welfare – generally asserting that Roma don't want to work and just go begging.<sup>103</sup>
- 12.7% - related to way of life, culture – stating that Roma are unadaptable, unchangeable, respect only their own rules...;
- 7.2% related to personal characteristics – referring to Romani persons as dirty, ugly, stupid and untrustworthy.

### CONCLUSIONS

Conclusions regarding online hate speech targeting Roma need to be set within the wider context of the continued rise of hate speech and 'communicative aggression' in Serbian public discourse and amplified in traditional and online media. This situation is aggravated by the dysfunctional system of media regulation, with a weak Press Council and social media operators who neither prevent nor remove hate speech. As ECRI concluded:

*"Many offences are not reported to the police and the police are not always open to receiving complaints, in particular from LGBTI persons and Roma. The application of the legislation against hate speech and violent hate crime is inefficient and there is no decisive action against the activities of racist, homo- and transphobic hooligan groups."<sup>104</sup>*

Ivana Krstić found that an increasing number of media outlets do not abide by professional principles, codes of ethics, and the language of tolerance. The daily, intensive use of aggressive and disturbing terminology used by the media, has desensitised citizens to *"the language of aggression to such an extent that once inappropriate words in public space or expressions used only in exceptional situations have become commonplace—part of the media, but also everyday vocabulary."<sup>105</sup>*

Research conducted by the Belgrade Centre for Human Rights concluded that events that polarise the public are the most common trigger of hate speech, especially towards Roma and LGBTIQ+ people, with an increase in online activity characterised by harsh aggressive language, conspiracy theories, insults, and a full lexicon of hateful and racist abuse.

As shown by the data produced from this research, most online hate speech targeting Roma surfaces on Facebook, which is in turn the most responsive social media platform in terms of response rates to reports and sending regular notifications regarding the status of reported content. Almost two-thirds of content reported to Facebook was removed. By contrast, YouTube in Serbia was found by the researchers to be 'totally unresponsive'.

<sup>103</sup> *Ibid.*

<sup>104</sup> Council of Europe, *ECRI REPORT ON SERBLA (fifth monitoring cycle)*. Adopted on 22 March 2017. Available at: <https://rm.coe.int/third-report-on-serbia/16808b5bf4>.

<sup>105</sup> Ivana Krstić, *Report on the Use of Hate Speech in Serbian Media*, Council of Europe, April 2021. Available at: <https://rm.coe.int/hf25-hate-speech-serbian-media-eng/1680a2278e>.

## FINDINGS

The research also revealed a lack of mechanisms for reporting hate speech on online news portals; either there was no option for reporting or there was an absence of any moderation after reporting in the comment section. This chimes not just with wider concerns about professional principles and codes of ethics discussed above, but also with the fact that media outlets and their online offshoots, owned or supported by the government, “*systematically deliver content to citizens that ... spreads disinformation and incites hatred.*” The EU Serbia 2020 Report similarly noted that: “*hate speech and discriminatory terminology are often used and tolerated in the media and are rarely tackled by regulatory authorities or prosecutors.*”<sup>106</sup>

One general observation from the volunteers’ monitoring is that there is a lack of consistency when it comes to social media networks removing reported hate content. While messages directly inciting racial hatred or violence against Roma were removed by Facebook, case analysis showed that in cases of reported content where racial slurs are coded and hate content is implicit rather than explicit, Facebook failed to recognise them as violating community standards and such content remained online. As Krstić noted, in recent years there have been more and more cases where explicit hate speech has been replaced by speech that is essentially hate speech, but is not recognisable as it at first sight. The most recent case concerns comments on social networks using the words ‘polar bears’ to mean Roma.<sup>107</sup>

It can be concluded that the combination of the failure to recognise implicit or coded hate speech, and the failure in some media to have any effective moderation of content or commentary, as identified by the Serbian research team, poses not just a threat to the security of Roma and other targeted groups, but also directly undermines the stability of democratic values in society. As Balkan Insight noted, social media companies should make sure that their content moderation processes – which combine the use of algorithms with the intervention of human moderators – are actually fit for purpose and capable of understanding the various dimensions of the local context. Failures to recognise and remove coded content can result in increased polarisation or even outbreaks of real-life violence, while weak or negligible content moderation can transform online platforms into “*hotbeds of disinformation, hate speech, and discrimination*”; a development Balkan Insight describes as “*especially concerning in post-conflict countries, where tensions between groups can erupt into violence.*”<sup>108</sup>

## RECOMMENDATIONS

The Serbian authorities should:

- Desist from the use of ethnic slurs, aggressive and hate speech against Roma, and promptly condemn any such incidents, online or offline; take appropriate and dissuasive action against political representatives or public officials who use or disseminate hate speech;
- Ensure that its laws criminalise incitement to racial hatred, whether or not it incites violence; and strengthen measures to ensure that racist hate speech, online and offline, is effectively identified, investigated, and punished. It should also ensure that prosecutors and judges receive intensive training on freedom of expression and hate speech, and that they are brought up to speed on pertinent EctHR jurisprudence;

<sup>106</sup> *Ibid.*

<sup>107</sup> *Ibid.*

<sup>108</sup> Roberta Taveri and Pierre François Docquir, ‘Online Hate Speech Remains Unmoderated in Balkans’, *BIRN* 21, June 2022. Available at: <https://balkaninsight.com/2022/06/21/online-hate-speech-remains-unmoderated-in-balkans/>.

- Combat the proliferation of racism and hate speech on the Internet; block overtly racist websites that incite violence and discrimination against Roma and other visible minorities; and require social media networks and news websites that feature online comments to monitor and effectively moderate their sites to remove hate speech;
- Initiate intensive and recurring training for journalists as a first response to what ECRI described as the “*frequent, serious breaches of the Code of Ethics*”; provide guidelines on hate speech that feature the relevant jurisprudence of the EctHR and practical examples to ensure journalists are fully aware of their public responsibilities;
- Ensure that the REM, the Press Council, and the Commissioner for the Protection of Equality have the resources and capacity to take up all cases of hate speech in the media, and can impose effective, proportionate and dissuasive sanctions on editors and journalists that produce and proliferate hate speech online;
- Regulate hate speech on the Internet in a comprehensive manner, in a way that delineates the particular responsibilities of the authors of hate speech and their editors, internet service providers, website forum hosts, social media platforms, and content moderators;
- Invest in preventative measures and step-up efforts to inform and sensitise the public about racist hate speech, discrimination, and antigypsyism, and work with civil society to actively promote inclusive democratic values and interethnic tolerance and understanding. Ensure that Roma and members of other groups targeted by hate speech are fully aware of, and can access, the relevant complaint mechanisms.

Social media networks should:

- Put rules and community standards in place that prohibit hate speech and commit sufficient resources to ensure that effective systems and staff are in place to promptly review content that is reported to violate these standards.
- Continuously review and revise their terms of service, rules or community standards to include a more precise definition of hate speech as prohibited content, and to prohibit users from posting content inciting violence or hatred.
- Make sure to take account of the specificities of antigypsyism; and see that community standards are fully applied to sanction online hate speech against Roma and other protected groups.
- Ensure that human moderators are native Serbian speakers familiar with anti-Roma narratives prevalent in Serbia, and adept at recognising implicit as well as explicit anti-Roma hate speech. Hold regular and frequent trainings, and provide coaching and support for their teams of content reviewers.
- Machine learning systems at social networks should be more sensitive at recognising ethnic slurs referring to Romani communities in Serbian.
- Intensify the work with civil society to deliver best practice training on countering hateful rhetoric and prejudice, and increase the scale of their proactive outreach to NGOs to help them deliver effective counter speech campaigns.
- YouTube needs to urgently improve its response rates in reviewing flagged material and removing hate speech content in Serbia. It should engage with civil society to increase familiarity among activists with the YouTube Trusted Flagger program, and explore how best to collaborate in taking down hate content.

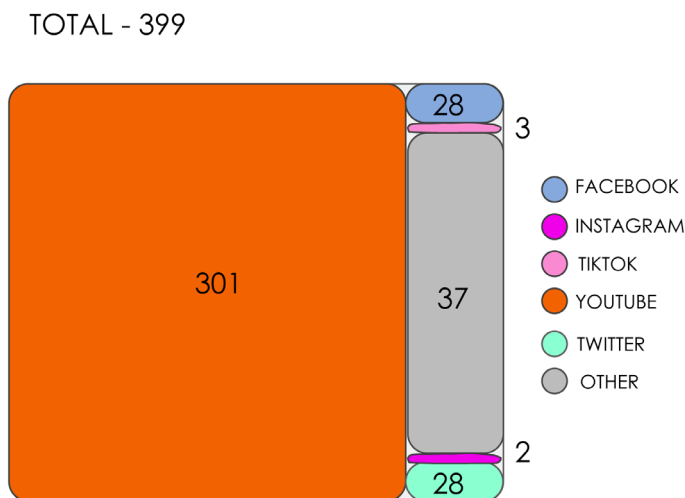
## Turkey

The Turkish team was engaged in the process of finding and reporting online hate speech content targeting Roma from November 2020 until June 2021.

Key words used for searching for anti-Romani narratives were: #Çingene- meaning Gypsy, and #Roman meaning Roma.

The Turkish team collected 399 examples of reported online hate speech targeting Roma. In terms of the type of reported content, comments were the most represented (353 cases), followed by posts (44 examples), one page, and one personal profile.

The majority of the examples were identified on YouTube (301 examples) and these were usually comments on videos. The number of examples from other sources is presented in the table below:



'Other' includes sources such as "Ekşi Sözlük: Sour Dictionary" and some online news portals.

Since the vast majority of the data is from YouTube, the Turkish national coordinator Serkan Baysak<sup>109</sup> was asked if YouTube is the network where hate speech targeting Roma is most commonly found. Baysak replied that they reported more than the 399 examples shown in the table across a number of sources, however YouTube is overrepresented as more examples taken from there were properly recorded and able to be coded as compared to examples from other networks.

### RELATION BETWEEN THE SOURCE AND STATUS

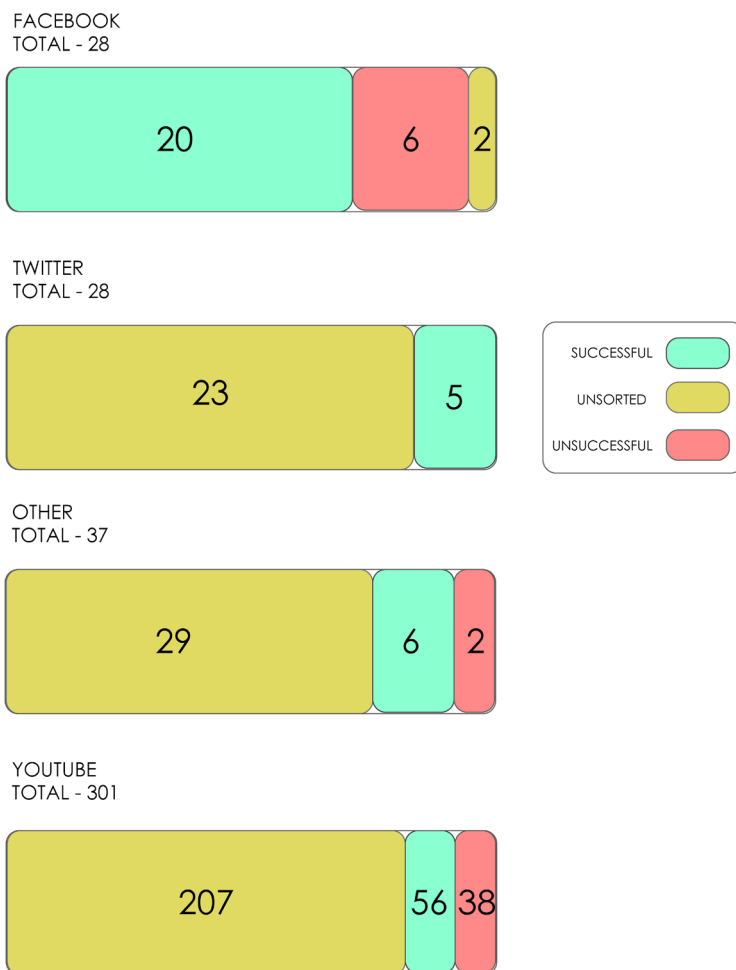
The majority of examples of reported content are unsorted (71.18%), meaning that information regarding whether the content was removed or not is missing.

<sup>109</sup> Interview with Serkan Baysak, October 2021.



Some data from the unsorted category may in fact have been successfully removed, but we do not have that information, related Baysak.<sup>110</sup> He added that Facebook normally sends notifications regarding the status of reported content, so the majority of unclassified data from Facebook (20 cases) is likely not because Facebook did not send a notification, but because the volunteers did not record whether the report was successful or not.

In the table below, data related to the status of reported content is shown for YouTube, Facebook, Twitter, and Other. There is a significant percentage of unsorted reports from each source so conclusions about the efficiency of the listed platforms in removing hate speech content cannot be accurately drawn. However, if the unsorted cases are excluded from analysis, the results show that for Facebook there were six successful reports and two unsuccessful. For YouTube there were 56 successful reports and 38 unsuccessful, and for Other six successful, and two unsuccessful. For Twitter, all five reports were successful.



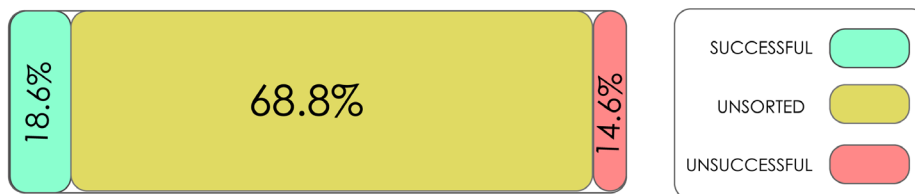
<sup>110</sup> *Ibid.*

## FINDINGS

Social media networks and platforms do apply censorship on content violating community standards, however more accurate results are needed in order to effectively analyse their efficiency and consistency.

Taking the results from YouTube as the main source of reported content, not taking into account the unsorted reports, it can be seen that there is no significant difference between successful (56 cases or 18.6%) and unsuccessful reports (38 cases or 12.6%).

TOTAL - 301

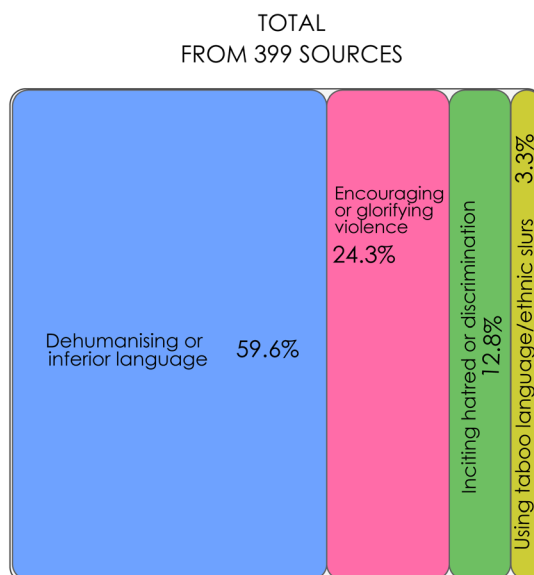


However, Baysak<sup>111</sup> said:

*“There were many serious hate speech examples on YouTube, but only some of the examples were considered as hate speech and removed. From this we can say that the policy of YouTube is almost zero.”* He also mentioned that Facebook was the most responsive platform.

## ANALYSIS RELATED TO HATE SPEECH DEFINITION

In the Turkish sample, almost 60% of reported content contained elements of ‘Dehumanising or inferior language’ as the main category of hate speech. ‘Encouraging or glorifying violence’ was found in 24.3% of the reports.



111 *Ibid.*

## CHALLENGING DIGITAL ANTIGYPSYISM: ALBANIA, SERBIA, TURKEY, AND UKRAINE

The table below shows that ‘Dehumanising or inferior language’ is the most frequently appearing category of the project’s hate speech definition found in all sources (with the exception of Instagram with only 2 reports in sum):

FACEBOOK - 28



INSTAGRAM - 2



TIKTOK - 3



YOUTUBE - 301



TWITTER - 28

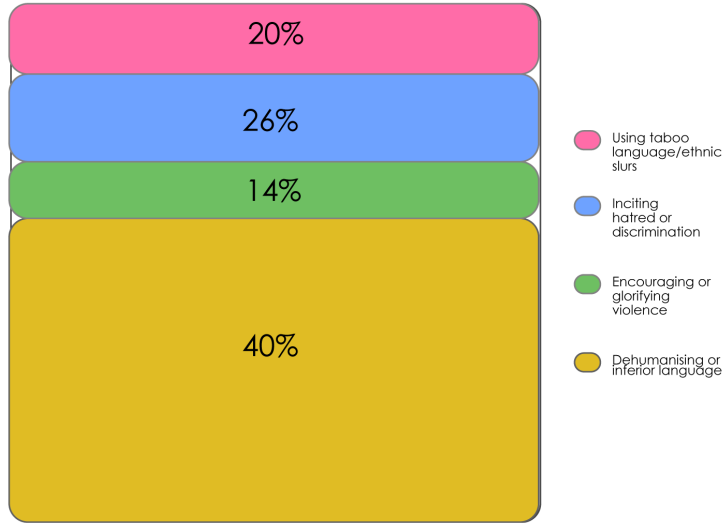


OTHER - 37



If all elements of the project’s hate speech definition found in the collected sample are taken into account, ‘Dehumanising or inferior language’ still dominates with 40%. The presence of ‘Using taboo language/ethnic slurs’ becomes more visible in this case:

TOTAL - 689

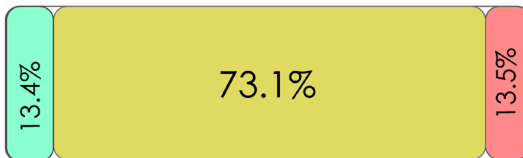


**HATE SPEECH DEFINITION AND STATUS**

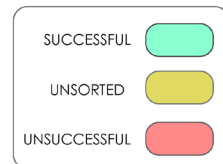
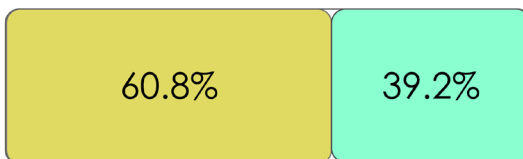
The Turkish data was analysed in terms of the type of content recognised by platforms as hate speech, and which type of content the platform took steps to remove.

Out of all the successful reports in the Turkish sample, 52% fall into ‘Encouraging or glorifying violence’ as the main category of the project’s hate speech definition (percentage related to status) and there are no unsuccessful reports in this category. This indicates that social media networks do react to this type of hate speech content.

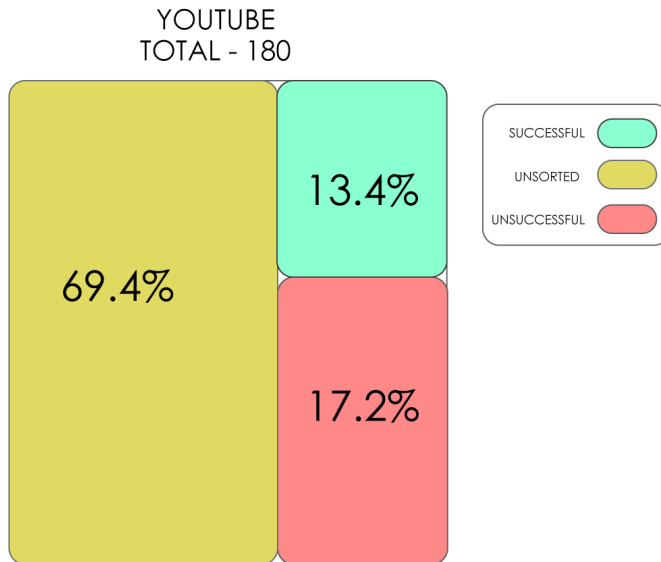
Dehumanising or inferior language  
TOTAL - 238



Encouraging or glorifying violence  
TOTAL - 97



In the Turkish sample, the majority of reports were designated as containing elements of ‘Dehumanising or inferior language’ as the main category of the project’s hate speech definition (238 samples or 59.7%). Not counting the unsorted reports, the same amount of reported content was removed by moderators (13.4%) as was left on the site (13.5%). A similar result can be seen if the data is filtered by YouTube as the source:



This indicates that there is no consistency in removing content with ‘Dehumanising or inferior language’ referring to Roma. In some cases this content is recognised as hate speech and in others it is not.

### TYPICAL ONLINE ANTI-ROMANI NARRATIVE

As demonstrated, examples of reported content containing ‘Dehumanising or inferior language’ dominate the Turkish sample. The example below is typical:

E Ebru Yazici • 9 ay önce

Osmanlı da roman nüfusu belgelerde buçuk olarak geçer yani osmanlı da 72.5 millet vardır 🤔o buçuk romanlardır 🤔🤔

👍 5   🗨️ 1

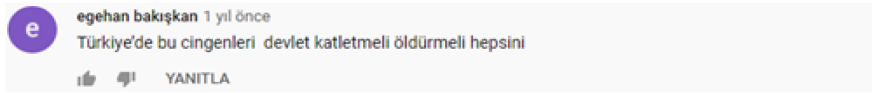
**Translation:**

There are 72.5 nations in the Ottoman Empire because in documents the Roma population is named as a half nation.

Serkan Baysak<sup>112</sup>, volunteer coordinator for the Turkish group, explained that the term “Çingene” which is an ethnic slur and a politically incorrect way to refer to Romani people, is commonly used and overrepresented in both the sample in comments on social networks and online media and in general. In the statistical analysis this is visible when examples containing all elements of the project’s hate speech definition are counted (138 cases or 20% of the whole sample).

<sup>112</sup> *Ibid.*

Additionally, explicit expressions of hatred and calls to exterminate all Roma (such as statements that all Roma should be killed) are very frequent, especially on YouTube. Baysak believes this is because people feel more free to express these attitudes there due to the absence of any prohibition concerning hate content, and the seeming lack of any notion of community standards on this platform in Turkey.<sup>113</sup>



### Translation:

The state should slaughter these gypsies in Turkey and kill them all.  
(Source: YouTube; Status: Successful)



### Translation:

Ethnicity? Kurd. In the Istanbul earthquake, the plunder by the Kurds and the gypsies will kill our people rather than the earthquake. MASSACRE. NEED TO BURN THE NEIGHBORHOODS. ALL ROMAN NEIGHBORHOODS SHOULD BE BURNED AT NIGHT.  
(Source: Twitter; Status: Successful)

The Turkish team analysed a sample of 399 cases of reported online hate speech based on the subject matter and key words in each case. All key words present in one example were counted for this analysis.

The most frequent narratives included the following:

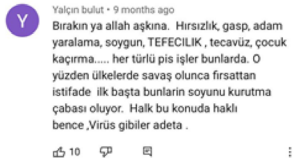
- Calls to exterminate or exile Roma from Turkey – counted 147 times in the entire sample;
- Personality characteristics such as Roma being shameless and dark-skinned, indigent, disrespectful, rude, not speaking correctly – counted 132 times;
- Describing Roma as violent, dangerous, as a threat to society, and ascribing them criminal behaviour – counted 123 times;
- Roma as thieves, with typical statements such as *stealing has become a lifestyle and profession for them* – counted 82 times;
- Roma as immoral and not religious – counted 70 times.

This analysis of the narratives favoured by anti-Roma racists online reflects wider societal prejudices against Roma. In addition to the social segregation and everyday racism Roma face from service providers, local authorities and the police, Roma are victims of mass evictions, gentrification clearances, and ‘lynching episodes’.<sup>114</sup>

<sup>113</sup> *Ibid.*

<sup>114</sup> Civil Rights Defenders, *Roma in Turkey: Discrimination, Exclusion Deep Poverty and Deprivation*, 2022. Available at: <https://crd.org/wp-content/uploads/2022/10/ROMAN-RAPORU-EN.pdf>.

As well as examples with clear violent messages, the Turkish national volunteers also selected a number of other examples depicting typical anti-Romani narratives they were encountering online. These cases were removed after being reported.



**Translation:**

Let it, God's sake. Stealing, extortion, injuring, robbery, usury, rape, kidnaps, all kinds of dirty works are done by them. Therefore, when there is war in countries, the first thing to take advantage of is the effort to kill all of them. I think the people are right about this, they are almost like viruses. (Source: YouTube; Status: Successful)



**Translation:**

Theft, fight, begging, filth, ignorance, taunts means Roma! I haven't met one good Roma! They know nothing but just to steal and steal. I can't trust any of them. It was a very optimistic video. I can't even call them human.

Prejudices related to religion are worth highlighting, since anti-Roma sentiments related to religion are not marked as significant in the samples of the other three target countries however they are present in a substantial number of examples (20) in the Turkish sample. Usually these sentiments profess a belief that Roma do not belong to a specific religion, which is seen as unusual and threatening to the majority population.



**Translation:**

Damn the Roma, because they are irreligious and do not fear Allah. They do not pray and are descendants of Pharaoh. Three gypsies stoned Muhammad. The only race cursed by Allah is the Roma. Let them die. They are parasites of the earth. (Source: Facebook; status: Successful)

It is possible that ‘Dehumanising or inferior language’ towards Roma, as a common theme in the Turkish sample, can be partially explained through these prejudices related to religion combined with ones related to accusing Roma of having no moral values (counted 50 times).

The incomplete anti-discriminatory legal framework, especially toward ethnic minorities, and weak institutions for human rights protection in Turkey perhaps contributes to a general perception of Roma as less worthy and inferior to the majority population, serving as justification of hatred towards this ethnic group.

## CONCLUSIONS

The proliferation of online anti-Roma hate speech occurs against a backdrop of democratic backsliding, where, as the European Commission put it, *“the rights of the most disadvantaged groups and people belonging to minorities need better protection”*; living conditions for Roma deteriorated severely, and gender-based violence, discrimination, and hate speech against minorities (especially LGBTIQ+ persons) are still a matter of serious concern.<sup>115</sup> The ruling AKP party enacted a state of emergency following the 2016 coup attempt which remained in effect until 2018. This allowed President Erdoğan to issue decrees without judicial oversight, including decrees that threatened freedom of expression online, which were used to block websites, shut down communication networks, and close civil society organisations and news outlets.

It is within this context that the combination of dehumanising narratives, abusive language, and ethnic slurs against Roma as revealed by the researchers surfaces online. The posts and comments they reported included incitement to commit violence, calls for extermination and exiling of Roma from Turkey, and narratives variously describing Roma not just as violent and dangerous thieves, but also as immoral and irreligious.

The forms and expressions favoured by anti-Roma racists online largely chimes with wider societal prejudices against Roma. The 2022 report by Civil Rights Defenders examined how Roma face everyday racism in their dealings with officialdom, service providers and local authorities, and how social segregation persists in informal daily life. The report highlighted issues around access to justice and the aggressive attitudes of the police and military, who perceive Roma as criminals and a security threat; a community to be profiled and controlled. In their daily life it is the security of Romani people that is under threat, and the report described how Roma are victims of mass evictions, gentrification clearances, and ‘lynching episodes’. Threats of violence against Roma have spilled over from the virtual into real world mob violence in the past, and the pogroms in Selendi in 2009 and Iznik in 2013, which forced hundreds of Romani people to flee and resulted in widespread damage to dwellings and property, were the most violent mob attacks against Roma in recent history.<sup>116</sup>

While hatred and incitement to hatred are prohibited under the Turkish Penal Code (TPC), as noted by the EU Commission; *“legislation on hate speech and its implementation need to be improved as it disregards hate speech against religions other than Islam”* and the legislation *“is not in line with the international standards.”*<sup>117</sup>

The regulation of Turkey’s media watchdog, the Radio and Television Authority (RTÜK), lacks clarity in terms of *“scope, definitions, licencing criteria and costs, and contains controversial provisions regarding jurisdiction and restricting access to online content.”* As regards hate speech, provisions of

115 European Commission - European Neighbourhood Policy and Enlargement Negotiations, *Türkiye Report 2022*, 12 October 2022. Available at: [https://neighbourhood-enlargement.ec.europa.eu/turkiye-report-2022\\_en](https://neighbourhood-enlargement.ec.europa.eu/turkiye-report-2022_en).

116 Beril Eski, *Roma In Turkey: Discrimination, Exclusion Deep Poverty and Deprivation*, Civil Rights Defenders, 2022. Available at: <https://crd.org/wp-content/uploads/2022/10/ROMAN-RAPORU-EN.pdf>.

117 Sinem Aydınlı and Brankica Petković, *Resilience: Civil Society for Media Free of Hate and Disinformation: Regulatory and self-regulatory framework against hate speech and disinformation in Turkey*, SEENPM, Tirana, Peace Institute, Ljubljana and Bianet, Istanbul, 25 November 2021. Available at: <https://seenpm.org/wp-content/uploads/2021/11/Resilience-Factsheet-Turkey-final-1.pdf>.



the Criminal Code are applied arbitrarily and the law is based on protecting those advantaged groups in favour of dominant ideology such as Islamic values, national unity and integrity, and Turkish family structure; and the expression ‘ethnic origin’ is not included in the Article regulating ‘hate and discrimination’ offences.

The attitude and approach of Criminal Courts’ judges to hate speech and hate crime does not favour the protection of minorities. It is rather the case that laws are enforced to limit freedom of expression, and criminal cases continue to be brought against, and convictions imposed on, journalists, human rights defenders, lawyers, writers, opposition politicians, students, artists, and social media users.<sup>118</sup>

In common with other countries, Facebook was found to be the most responsive platform, sending notifications regarding the status of the reported content despite inconsistency in its application of community standards. At the other end of the scale YouTube in Turkey was described by the research team as “very unresponsive” to such an extent that it was not possible to accurately track the status of reported cases, which included extreme threats of violence against Roma.

In terms of restrictions and prohibited content, what distinguished Turkey was the massive investment by the coercive apparatus of the state to stifle freedom of expression and any dissenting content critical of the regime or its purported ‘traditional values’. Under the pretext of regulating social media networks, in July 2020 the ruling AKP introduced a set of restrictive measures such as the obligation of each social network to have a representative present in Turkey responsible for the removal of illegal content and delivery of user information when requested, and the obligation to store Turkish users’ information in Turkey, in addition to a number of other measures aiming to restrain social media usage.

Government officials claimed this was to ensure that crimes committed on social media platforms will not go unpunished. But as EuroMed Rights noted, in a country where the regime leaves no room for divergent discourses, “most often, these laws are already implemented solely for the purpose of censorship. The existing legislative regulations are also used in order to limit freedom of expression and not only to punish perpetrators of hate speech and defamation.”<sup>119</sup>

## RECOMMENDATIONS

The Turkish government should:

- Amend current legislation to include a clear definition and a prohibition of direct expression of hate speech that is compliant with international conventions and recommendations as they relate to related to the elimination of hate speech based on ethnic identity; and introduce proportionate sanctions for incitement to hatred against ethnic minorities, including Roma.
- Through parliament, change the procedure in the laws regarding the participation of NGO monitoring and countering the hate speech and disinformation in media in the proceedings

<sup>118</sup> *Ibid.*

<sup>119</sup> EuroMed Rights, *Turkey’s Social Media Bill, Another Obstacle to Freedom of Expression*, 31 July 2020. Available at: <https://euromedrights.org/publication/turkeys-social-media-bill-another-obstacle-to-freedom-of-expression/>.

to support the victims of hate speech and disinformation, as Turkish law does not recognise the standing of NGOs to bring claims in support of victims of discrimination.

- Adopt codes of conduct prohibiting hate speech in general, and refrain from anti-Roma hate speech and the use of ethnic slurs such as '*Çingene*' to refer to Roma in particular. Encourage all political parties to do likewise.
- Take a proactive role to collaborate with media, educational institutions, and civil society in awareness-raising campaigns to challenge and change negative stereotypes and racist prejudice against Roma.
- Ensure judges and lawyers receive adequate training to ensure hate speech legislation is applied in compliance with European Court of Human Rights jurisprudence, and facilitate the establishment of a politically autonomous expert body to conduct media monitoring to render visible hate speech content wherever it surfaces.

Social media platforms should:

- Apply rules, community standards, and content moderation practices in Turkey that are commonplace across Europe. Pay particular attention to those provisions that prohibit hate speech, and as a matter of urgency, commit sufficient resources to ensure that effective systems and staff are in place to promptly review content that is reported as violating these standards.
- Machine learning systems at social networks must be improved, and human moderators need to become more adept at recognising ethnic slurs referring to Romani communities. Invest in continuous training, coaching, and support for content reviewers to ensure optimal and consistent responses to reported hate speech against Roma.
- Intensify the work with civil society to deliver best practice training on countering hateful rhetoric and prejudice, taking account of the specificities of antigypsyism, and increase the scale of their proactive outreach to NGOs, educators, and experts to help them deliver effective counter-speech campaigns.
- YouTube needs to urgently improve on its apparent zero-response rates in reviewing flagged material and removing hate speech content in Turkey. It should adopt the standards it has pledged to maintain inside the EU and apply them in Turkey. As a matter of urgency, YouTube should engage with civil society to increase familiarity among activists with the YouTube Trusted Flagger program, and explore how best to collaborate in taking down hate content.

## Ukraine

The Ukrainian team was engaged in the process of finding and reporting online hate speech content targeting Roma from December 2020 until May 2021. At the time of editing this four-country report, more than a year has passed since the full-scale invasion of Ukraine by Russian forces on 24 February 2022. Since the invasion, between 10,000 and 13,000 Ukrainian soldiers have been killed, and as of 6 November 2022, OHCHR had recorded 16,462 civilian casualties: 6,490 killed and 9,972 injured. The Russian onslaught also prompted the flight of more than three million people, spurring the largest refugee crisis in Europe since the Second World War.<sup>120</sup>

As the CSCE stated, “*Well-documented Russian bombings and missile strikes in Ukraine have decimated hospitals, schools, and apartment buildings ... The withdrawal of Russian troops from towns like Bucha, Chernihiv, and Sumy has revealed horrific scenes of civilian carnage, mass graves, and reports of rape and torture. Several world leaders have accused Russia of committing genocide against the people of Ukraine.*”<sup>121</sup>

In short, the context in which the following research and monitoring was carried out no longer exists.



Key words used for searching were ‘Roma’ and ‘Cigany’, with ‘Cigany’ representing an ethnic slur for the Romani community. As there are a large number of Russian speakers as well as Ukrainian speakers in Ukraine, volunteers used both the Ukrainian and Russian spellings of the slur ‘Cigany’: ‘Цигани’ in Ukrainian and ‘Цыгане’ in Russian.

The Ukrainian volunteers reported that on YouTube it was easy to find examples of anti-Romani hate speech using the word ‘Cigany’ in Russian or Ukrainian. On the other hand, negative content about Roma was quite difficult to find on Facebook by searching with the key words, except on the pages of far-right groups where the majority of the reported content was found.

The volunteers categorised three of the most searched sources in the following way:

- 1) social media pages of far-right groups;
- 2) neutral entertaining platforms;
- 3) “neutral” platforms with infiltrated right-wing and xenophobic followers. Examples of these are: “Kiev operative”, “Municipalna Varta”, “News of Izmail, Ivano-Frankivsk”. These types of platforms and groups portray neutrality but in reality, they spread hatred toward minorities in Ukraine, including Roma.<sup>122</sup>

<sup>120</sup> Office of the UN High Commissioner for Human Rights (OHCHR), *Ukraine: civilian casualty update 7 November 2022*. Available at: <https://www.ohchr.org/en/news/2022/11/ukraine-civilian-casualty-update-7-november-2022>.

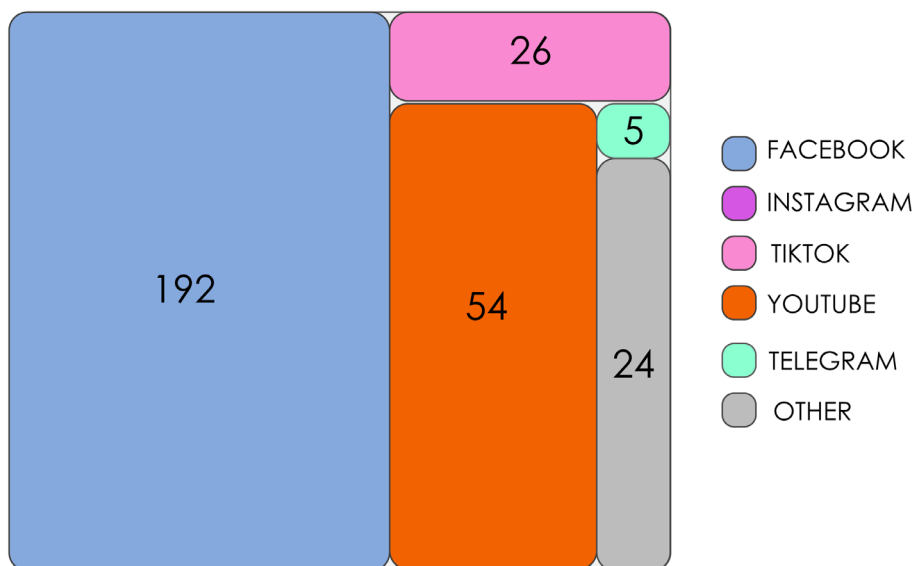
<sup>121</sup> Commission on Security and Cooperation in Europe (CSCE), *Ukraine: civilian casualty update 7 November 2022*. Available at: <https://www.csce.gov/international-impact/events/russian-war-crimes-ukraine>.

<sup>122</sup> Interview with Nataliia Tomenko, November 2021.

## FINDINGS

The Ukrainian volunteers collected 301 examples of reported online hate speech targeting Roma. Comments were the most frequently appearing type of reported content (188). Interestingly, compared to the results of the other national groups participating in this research, posts were represented in a significant amount (81 cases). Posts were usually identified and reported on Facebook. There were also 20 pages and 12 personal profiles reported. The volunteers estimate that approximately 70% of the reported content is in text format, and the remaining 30% in photo or video format. Most of the examples were identified on Facebook (192). The number of examples from other sources is presented in the table below:

### TOTAL - 301



The label “Other” refers to websites of the Ukrainian national and local media channels.

### EXPERIENCES WITH DIFFERENT SOCIAL MEDIA CHANNELS<sup>123</sup>



The Ukrainian Volunteer Coordinators found that **Facebook** usually sends notifications regarding the status of reported content. Hate speech content was found mostly on the pages of far-right groups, and it often was not removed from Facebook due to the fact that the content was coded. This is where the authors of the content use a coding language, where some of the letters in words are substituted with symbols such as: @ \$ \* - eg; ‘Cig@ny’. This decreases the chance that Facebook will remove the content. Members of far-right groups also use more evasive language; they do not write directly “*let’s kill Roma*” but instead use phrases such as “*Hitler knew what to do with them*”.

<sup>123</sup> All the information in this section comes from interviews with Nataliia Tomenko and Volodymyr Yakovenko, October 2021.



The Ukrainian team pointed out that **Telegram** is highly used among some groups for exchanging information and spreading hatred that can be harmful for Romani communities. The volunteers were not able to identify many reports from Telegram because this messaging application is mostly used for closed group messaging, hence why it is popular amongst far-right groups. However, there are also some open Telegram channels, such as Kyiv Operativ and Suganipartyl, which were important sources for this research.

On Telegram there are only a small number of reasons offered as to why you are reporting content, and there is no ‘hate speech’ option among them. The Ukrainian volunteers chose the ‘Other’ option, which would then further offer the ability to describe in written form the reason for reporting. After reporting, they would usually receive a notification from Telegram as follows: ‘Your information has been received by a moderator’. It also proved difficult to identify and report the creator/moderator of the channel in Telegram, another reason why far-right groups are prevalent on the app.

It was also not possible to search for key words on Telegram using an overall search function, because the content is shared in closed groups where you have to be accepted as a follower in order to see the content.



It proved easy to find plenty of anti-Roma comments below the videos on **YouTube**, but after reporting comments were usually not deleted.

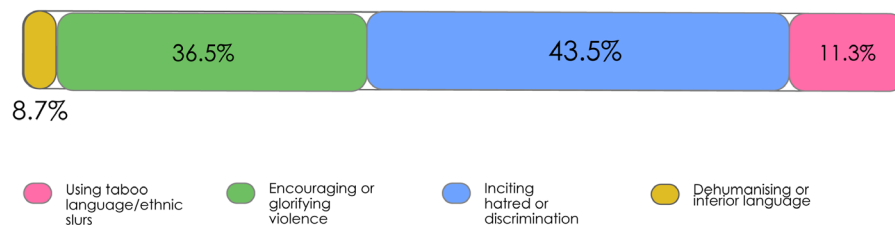


After reporting a video on **TikTok**, volunteers would usually receive the following response: “Thank you for your report, from now on, you won’t be seeing videos with similar content.” In order to see the status of the reported video, it is necessary to create another TikTok account or to try to see the status through another account.

The volunteers also noticed that by reporting content the video’s rating on TikTok goes down, and as a consequence the number of views also decreases. Additionally, if three videos from the same account are reported, the account automatically becomes banned for a period of time, but the videos are still not removed from TikTok.

When considering the main category of the project’s hate speech definition in the Ukrainian sample, it can be seen that examples which fall under ‘Inciting hatred or discrimination’ and ‘Encouraging or glorifying violence’ dominate.

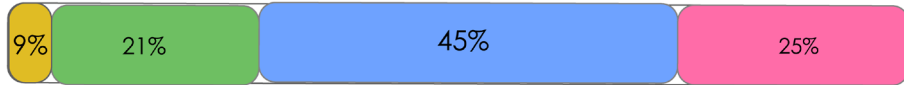
TOTAL - 301



## FINDINGS

When all elements of the hate speech definition are taken into account, the number of examples 'Using taboo language/ethnic slurs' increased significantly, more than doubling from 11.3% to 25%:

### TOTAL - 503



The majority of reported content from Facebook is categorised as containing elements of 'Encouraging or glorifying violence' (92 cases), followed by content 'Inciting hatred or discrimination' (66 cases). Examples from YouTube are mostly fall under the 'Inciting hatred or discrimination category' (39 out of 54 cases).

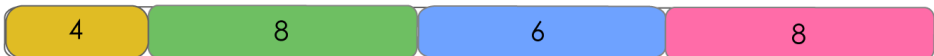
### FACEBOOK - 192



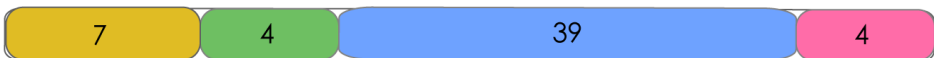
### TELEGRAM - 5



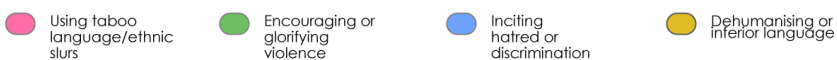
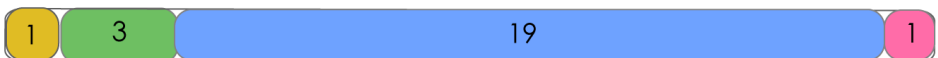
### TIKTOK - 26



### YOUTUBE - 54



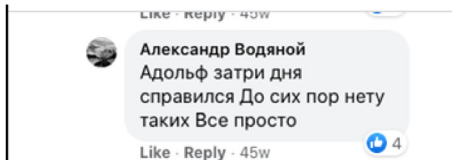
### OTHER - 24



Bearing in mind that the majority of cases from Facebook were found on the pages of far-right groups, it is not surprising to see so many cases labelled as 'Encouraging or glorifying violence'.

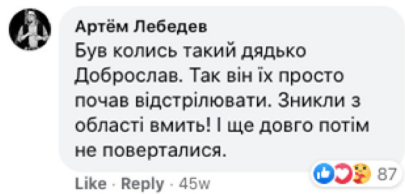
## TYPICAL ANTI-ROMANI NARRATIVES

The Ukrainian volunteers also identified examples which present a narrative that promotes violence towards Roma:

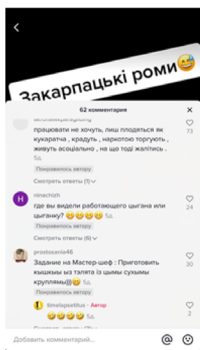


**Translation:**  
Adolf coped in three days!  
There are still no such (person as Hitler).  
Everything is simple.

**Translation:**  
There was once such an uncle Dobroslav (the person who conducted hate crimes against Roma). So he just started shooting them. They disappeared from the area in an instant! And they did not return for a long time.

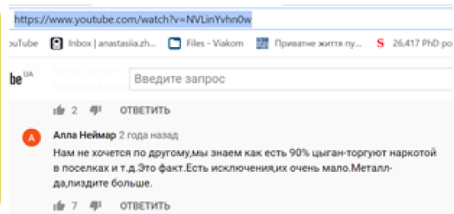


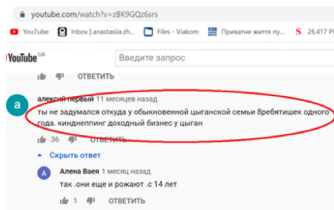
They also identified examples which illustrate common stereotypes and prejudices towards Roma that the volunteers regularly encountered online:



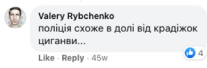
**Translation of the first two comments:**  
They don't want to work, they just breed like cockroaches, they steal, sell drugs, live asocially, so why do they complain then?  
Where did you see a cigan who works?

**Translation:**  
We know that 90% of cigans sell drugs in settlements, etc. This is the fact. There are exceptions, but there are just very few of them. You say they sell metal, "yeah", bullshit.





**Translation:**  
Probably you have not thought about where an ordinary gigan's family has 8 children with the same age. Kidnapping is a profitable business for cigans.



**Translation:**  
Police is probably in cooperation with cigans...

## ONLINE FAR-RIGHT ACTIVITIES SPOTTED IN THE MONITORED PERIOD

The Ukrainian volunteers highlighted the usage of Telegram in spreading far-right ideologies. Open Telegram channels worth mentioning for openly promoting hatred against Roma are Kyiv Operativ and SuganiPatryl.

Kyiv Operativ<sup>124</sup> - (157K subscribers)

Kyiv Operativ is a far-right news portal and has related social media channels on Facebook, YouTube, Telegram, Instagram, Viber, and Twitter. While on Facebook and Twitter they have a moderate following; 8,714 page likes on Facebook and more than 12000 follows, on Telegram they have 157k subscribers.<sup>125</sup> There is a significant difference in reach of the same post shared on the different social media networks; Kyiv Operativ can gather around 56 comments on a Facebook post but over 16k on Telegram.

Presenting themselves as ‘Guardians of Kiev’, their goal is allegedly to prevent crimes. They mostly post content related to non-white communities, where crimes or possible crimes are committed. It is particularly Roma who are depicted, and always negatively as thieves, beggars etc. Each of the posts related to Roma has a comment section containing numerous examples of hate speech.

A particularly dangerous trend the volunteers noticed in the channel of this group on Telegram is that subscribers post photographs of Romani people as well as map locations labelled as ‘dangerous spots’. These could be photographs of Romani people begging on the street or photographs of the Romani people with an indication of their location and descriptions such as ‘be careful at the metro station... there is a group’. Below such posts, there are comments regarding the ethnicity of the group (Roma) and this is usually followed with hate speech towards the whole Romani community. These cases became systematic and followers of the group send pictures into the channel of Romani people with their location.

<sup>124</sup> Kyiv Operative on Telegram. Available at: [https://t.me/KyivOperativ?fbclid=IwAR3jvPV1iOlgwM0LyW MU\\_suDEB1k74QuRfP3NK3DSvTxdZXH69o0Vu88jGM](https://t.me/KyivOperativ?fbclid=IwAR3jvPV1iOlgwM0LyW MU_suDEB1k74QuRfP3NK3DSvTxdZXH69o0Vu88jGM) (accessed November 2021).

<sup>125</sup> Data from November 2021.



Sugani patryl “Gypsy patrol”<sup>126</sup> - (515 subscribers)

This group operates in a similar way to the previous group, however they are solely focused on Roma and they have fewer subscribers. They also indicate locations where Roma have been seen, to “prevent” crimes since Roma are considered to be thieves.

This trend of mapping locations is illustrated by the below screenshot taken from Facebook which shows mapping of “Gypsies crimes” in Kyiv.

**Translation:**

Kyiv is the capital of robbers. This map shows several thousand cygan crimes in the last FIVE months. It's impressive! Tens of millions of hryvnias were stolen. Thousands of broken lives of victims.

This is what Kyiv looks like through the eyes of a cygan.

Five months of our analytical work. There is a database of a half thousand robbers from dozens of gangs. These are just the crimes that the victims have informed us about in this short time. And no unwashed abomination was put into prison.

By reposting (this post) you will save your daughter's mother's or neighbour's wallet.

The author of the post is a member of the “Municipal Varta”, Volodymyr Irlandets. Municipal Varta is an NGO that “protects society from dangerous people (mostly members of minorities)”: <https://www.facebook.com/MunicipalnaVarta/>.

## IN LIEU OF A CONCLUSION

Nine months after the full-scale invasion of Ukraine by Russia on 24 February 2022, the European Parliament declared Russia to be a state sponsor of terrorism. The resolution, passed on 23 November 2022, condemned “the deliberate attacks and atrocities committed by Russian forces and their proxies against civilians in Ukraine, the destruction of civilian infrastructure and other serious violations of international and humanitarian law (which) amount to acts of terror and constitute war crimes.”<sup>127</sup>

<sup>126</sup> Sugani patryl on Telegram. Available at: [https://t.me/sugani\\_patryl?fbclid=IwAR2wNwfE-GijnSgsYD1t48I-Hix0A7c7f2nPIqKgrNmOwKGjSsL0mEn4pMBI](https://t.me/sugani_patryl?fbclid=IwAR2wNwfE-GijnSgsYD1t48I-Hix0A7c7f2nPIqKgrNmOwKGjSsL0mEn4pMBI) (accessed November 2021).

<sup>127</sup> Delegation of the European Union to Ukraine, *European Parliament declares Russia to be a state sponsor of terrorism*, 23 November 2022. Available at: [https://www.ecas.europa.eu/delegations/ukraine/european-parliament-declares-russia-be-state-sponsor-terrorism\\_en](https://www.ecas.europa.eu/delegations/ukraine/european-parliament-declares-russia-be-state-sponsor-terrorism_en).

In this context, forging a conclusion or issuing a set of recommendations to the Ukrainian government on challenging online hate speech would make little sense while that government is engaged in combating massive waves of Russian airstrikes, aimed at destroying its energy infrastructure. The following is therefore merely a summary of the findings of research conducted before the war.



The majority of reported content from the Ukrainian sample came from Facebook (192 cases) followed by YouTube (54). The most represented examples were reported comments, however, compared to other national data, the Ukrainian team also identified a significant number of reported posts – 81. These posts were mostly from Facebook.

Key words used for searching were ‘*Roma*’ and ‘*Cigany*’, the latter representing an ethnic slur referring to the Romani community. The Ukrainian team also monitored the Facebook pages of far-right groups, where a trend of coding words and messages was discovered, decreasing the chances that the hate speech will be identified and censored.

Examples of hate speech which fell under the ‘Inciting hatred or discrimination’ and ‘Encouraging or glorifying violence’ main categories of the project’s hate speech definition dominated the sample. The number of examples which were seen to be ‘Using taboo language/ethnic slurs’ increased when all elements of the project’s hate speech definition were counted. Examples labelled as ‘Encouraging or glorifying violence’ were mostly identified on Facebook, which is not surprising when taking into consideration that the volunteers found a large number of the examples on far-right groups’ Facebook pages.

The Ukrainian volunteers also highlighted the prevalence of the messaging app Telegram among far-right groups, and groups allegedly active in preventing crimes but openly expressing hatred towards Roma and other minorities, such as Kiev Operativ. These groups post locations where Roma are seen and posts related to Roma incite numerous examples of hate speech. A trend of mapping locations with ‘Gypsy crimes’ was also detected on Facebook, by the followers of far-right pages and groups.

Qualitative data indicates that the typical anti-Romani narratives in Ukraine are concentrated around the perception of Roma as thieves and dangerous. This kind of narrative is not sanctioned; on the contrary, it serves as justification to formal and informal groups to proclaim their racist attitudes. Besides Kiev Operativ, other groups with a racist agenda identified in this research were: Municipalna Varta, News of Izmail, Ivano-Frankivsk, and Sugani partyl.

## General Conclusions

In the four countries, researchers found online hate speech and content disparaging Romani people ranged from ridicule to direct neo-Nazi calls to commit acts of racist violence against Roma. Beyond the immediate threat to the safety and well-being of those targeted by haters, the cumulative impact of this relentless online hate speech is to further normalise antigypsyism in the real world.

The concerns of the researchers were shared by a European Parliament study which found that rising hate speech ‘poisons society’ and surfaces at the highest levels, and both political actors and citizens ‘express their thoughts without inhibition’ on social media.<sup>128</sup> For its part, the European Commission described the “sharp rise in hate speech and hate crime across Europe – offline and online” as a particularly serious and worrying phenomenon, and proposed to extend the list of EU crimes to hate speech and hate crime.<sup>129</sup>

Despite the very different national contexts, some common themes emerged across the four countries surveyed. While removal rates of reported hate speech content were better on Facebook than other platforms, the monitors found there was little consistency and lots of failures to recognise thinly-coded hate content or racist dog-whistles, as well as cases of Facebook moderators simply deciding that ethnic slurs did not violate community standards.

As mentioned earlier, the volunteers were driven by a shared conviction that anti-Roma hate speech had been overlooked for too long, and a collective desire *to do something about it*, to develop practical and effective responses to counter online hatred and its consequences. While each section contains country-specific recommendations, the volunteers found common ground in formulating recommendations about action that needs to be taken by the authorities and social media platforms to stem the flow of online hate against Roma and other racialised communities across Europe.

This included calls for clearer definitions of hate speech, prohibitions of direct incitement, and proportionate sanctions for hate speech that incites violence against racialised minorities, including Roma. In each of the countries the authorities were also urged to take a proactive role in collaborating with media, educational institutions, and civil society in awareness-raising campaigns to challenge and change negative stereotypes and racist prejudice against Roma.

For their part, social media platforms were called upon to honour the commitments they have publicly made, and for the most part fail to honour, especially when it comes to

<sup>128</sup> European Parliament Committee on Civil Liberties, Justice and Home Affairs (LIBE Committee), *Hate speech and hate crime in the EU and the evaluation of online content regulation approaches*, European Parliament, July 2020. Available at: [https://www.europarl.europa.eu/RegData/etudes/STUD/2020/655135/IPOL\\_STU\(2020\)655135\\_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/STUD/2020/655135/IPOL_STU(2020)655135_EN.pdf).

<sup>129</sup> European Commission Press Release, *The Commission proposes to extend the list of ‘EU crimes’ to hate speech and hate crime*, 9 December 2021. Available at: [https://ec.europa.eu/commission/presscorner/detail/en/ip\\_21\\_6561?fbclid=IwAR1F5ebLg-11b1tVi\\_-O7WMfRZnE4g4Ga35Ozfu1vBq-PuXTCjXDghoE8I](https://ec.europa.eu/commission/presscorner/detail/en/ip_21_6561?fbclid=IwAR1F5ebLg-11b1tVi_-O7WMfRZnE4g4Ga35Ozfu1vBq-PuXTCjXDghoE8I).

## GENERAL CONCLUSIONS

the prompt removal of hate content and disinformation targeting Romani communities. The monitors in each country stressed the need for sufficient resources to ensure that effective systems and staff are in place to promptly review content that is reported as violating community standards, and to provide continuous training and support for content reviewers to ensure optimal and consistent responses to reported hate speech against Roma. The social media platforms were also urged to step up, intensify their work, and be far more proactive in their outreach with civil society to NGOs, educators, and experts to build common capacities to campaign and counter hate speech and racist prejudice in a manner that takes into account the specificities of antigypsyism.

Perhaps the most important message from these participants in the ERRC's volunteer-led project *Challenging Digital Antigypsyism* is the importance of agency. By forming digital activist communities, by taking action to monitor, record, and report examples of anti-Roma hate speech, and by holding political authorities and social media platforms publicly to account, these teams of volunteers have performed an exemplary civic duty in defending democracy. By their actions, they offer a vital corrective to the jaded acceptance of dehumanising narratives, abusive language, and ethnic slurs against Roma as an inevitable and pervasive feature of social media, and a standing rebuke to the notion that antigypsyism can ever be Europe's 'last acceptable form of racism'.



errc