BULGARIAN NATIONAL REPORT

Project: Together Against Antigypsyism Online (TAAO)

Programme: Citizens, Equality, Rights and Values Programme (CERV-2023-EQUAL)

Project number: ----------

Duration: March 2024 – February 2026

Project Leader: Amaro Drom e.V. (Germany)

National Partner: Integro Association, Bulgaria

Countries involved: Belgium, Bulgaria, Czech Republic, Hungary, Slovakia, Romania

Authors: K. Makaveev
V. Ibryam

0. INTRODUCTION / FOREWORD AND BACKGROUND – THE ROLE OF INTEGRO ASSOCIATION

Association Integro is one of the most prominent Roma non-governmental organizations in Bulgaria. Founded in 2002, it brings together a team of committed professionals working to address the deeply rooted challenges faced by the Roma community. These challenges—including stereotypes, prejudice, social exclusion, segregation, discrimination, and a specific form of racism known as antigypsyism—persist in Bulgaria and across Europe.

The social marginalization of Roma people is reinforced by negative public attitudes, which are often reproduced and amplified through media narratives and public discourse. In the early 2000s, when print media, radio, and television were the primary sources of information, Integro systematically analyzed the representation of Roma communities in public media. The organization has consistently stressed that freedom of speech cannot be used to justify violations of human rights, and that protecting the dignity of every individual is a cornerstone of democratic society.

Believing that sustainable change begins with young people, Integro implemented the European REACT campaign of the ERGO Network in Bulgaria in 2010, under the slogan "Thank You, Mayor!". This initiative launched the first systematic study of Roma representation in print media, conducted over a 12-month period by eight young Roma professionals and students. The findings, published in 2011, highlighted persistent negative and stereotypical portrayals of Roma communities and laid the foundation for subsequent advocacy campaigns against hate speech in media.

In the following years, Integro expanded its monitoring to radio and television. Within the framework of the project *Curbing Anti-Gypsyism from Local to European Level*, funded by the Open Society Foundation – Budapest, the report *The Image of Roma in Six Electronic Media* was published in 2015. These experiences demonstrated that media narratives could shift positively when young people actively engage in defending human rights.

With the advent of digitalization, hate speech increasingly migrated to online platforms, spreading rapidly and reinforcing discriminatory attitudes. In response, Integro joined the *Peer Education to Counter Antigypsyist Online Hate Speech* (PECAO) project, coordinated by ERGO Network, in 2020. Building on PECAO's foundations, the current project *Together Against Anti-Gypsyism Online* (TAAO) unites young people from six European countries who are committed to actively confronting online hate.

In Bulgaria, between October 2024 and September 2025, a national team of young Roma interns and professionals monitored 528 instances of online hate speech. All participants completed a two-week online training led by Assoc. Prof. Dr. Habil. Ileana Rotaru (West University of Timișoara), where they learned to identify, classify, and analyze hate content in accordance with the project's ethical and methodological standards.

Integro's monitoring team comprised five young professionals, supervised by Kamen Makaveev (Work Package 4 Coordinator) and Vergil Ibryam (Work Package 2 Monitoring Lead). The team

held regular meetings, provided supervision and feedback, and summarized their findings in quarterly analytical reports.

Early detection and mitigation of online hate speech are critical for fostering a safe digital environment for younger generations. This report provides both quantitative and qualitative analyses of online content targeting the Roma community in Bulgaria, highlighting the proactive efforts of young monitors in countering online hate and promoting social inclusion.

0.1. SUMMARY OF THE TAAO PROJECT – BULGARIA

The *Together Against Anti-Gypsyism Online* (TAAO) project, funded under the Citizens, Equality, Rights and Values (CERV-2023-EQUAL) programme, tackles one of the most persistent and under-documented forms of racism in Europe: antigypsyism. The project aims to examine how antigypsy hate speech spreads across digital platforms, how it is manifested among users and media outlets, and how online platforms respond to such content.

TAAO is coordinated by Amaro Drom e.V. (Germany) and implemented in six EU member states: Bulgaria, Czechia, Germany, Hungary, Romania, and Slovakia. The project aligns with EU priorities to safeguard fundamental rights and values by addressing hate speech and hate crimes, with a specific focus on combating antigypsyism in the online environment.

Within the project, national teams of at least five trained monitors in each country systematically observed, recorded, and analyzed cases of antigypsy hate speech following a unified European methodology. In Bulgaria, the team identified and classified 528 cases between October 2024 and September 2025, covering social media platforms, news portals, and other digital sources. The collected data were organized by thematic focus, type of discourse, level of aggressiveness, and platform response.

The Bulgarian contribution comprises four synthesized reports, corresponding to the four monitoring periods, with a total of 528 completed Monitoring Tools (MT) – the project's core research instrument. The structure and application of the MT are detailed in Appendix 1.

The analysis highlights the resilience of antigypsy narratives, which perpetuate negative stereotypes, undermine Roma identity, and normalize social exclusion. Many online posts employ coded language, irony, or humor to obscure discriminatory messages, demonstrating how hate speech can remain subtle, socially tolerated, and largely unnoticed in digital environments.

Findings also reveal significant disparities in platform responses, ranging from immediate content removal to complete inaction. By providing structured national data alongside comparative analyses, TAAO contributes to a nuanced understanding of online hate speech dynamics and the ways in which antigypsyism adapts to evolving forms of digital communication.

These insights hold practical relevance for policymakers, educators, and civil society actors—both Roma and non-Roma—by informing the development of more effective strategies for preventing, reporting, and countering online hate speech.

This national report was prepared by Kamen Makaveev with the support of the Bulgarian monitoring team and Assoc. Prof. Dr. Habil. Ileana Rotaru (West University of Timișoara). Special acknowledgment goes to the Bulgarian team—Ahmed Hasanov, Vergil Ibryam, Gamze Maksimova, Diyan Dankov, and Maria Hristova—for their professionalism and dedication. Overall project coordination and quality management were ensured by Liliya Makaveeva, Executive Director of Association Integro.

1. INTRODUCTION

1.1 Background and Context

The Bulgarian national report is part of a multinational study conducted in six European Union member states—Germany, Bulgaria, Czechia, Hungary, Romania, and Slovakia—within the framework of the *Together Against Antigypsyism Online* (TAAO) project. Coordinated by Amaro Drom e.V. (Germany) and implemented with the support of national partners, the project is funded under the Citizens, Equality, Rights and Values (CERV-2023-EQUAL) programme and runs from March 2024 to February 2026.

The primary objective of TAAO is to counter online hate speech targeting the Roma community through coordinated monitoring, public awareness-raising, and advocacy actions at both national and European levels. The project seeks to foster civic responsibility in digital spaces by promoting awareness, empathy, and respect for human dignity.

In Bulgaria, monitoring activities were implemented by Association Integro between October 2024 and September 2025. The national team consisted of five trained monitors who collected and analyzed 528 instances of antigypsy hate speech across social media platforms, news portals, tabloids, and other online sources.

Prior to the start of monitoring, all participants completed a two-week online training led by Assoc. Prof. Dr. Habil. Ileana Rotaru, which focused on research methodology, data collection processes, and ethical standards for online monitoring.

Throughout the project, monthly online coordination meetings were held with representatives from all partner countries to exchange updates on progress, harmonize methodologies, and share experiences. In cases where direct participation was not possible, updates were provided via recorded sessions or individual discussions within dedicated internal communication channels for national teams and work package coordinators.

1.2 Significance of the Study

In recent years, the use of the internet, social media, and digital professional and educational platforms has grown exponentially. While these developments offer significant opportunities, they also expose individuals—including children and youth—to various forms of online hate speech, whether as witnesses, participants, or victims. In this context, the Roma community remains one

of the most affected groups, with manifestations of discrimination and stereotyping potentially leading to serious social and psychological consequences.

The TAAO project (*Together Against Antigypsyism Online*) builds on the efforts of the PECAO project (*Peer Education to Counter Antigypsyist Hate Speech Online*) and focuses on systematically tracking and countering antigypsy hate speech online. The Bulgarian study draws on previous European research and aims to:

- Provide comprehensive data on the prevalence, impact, authors, and recipients of antigypsy hate speech, as well as societal and platform responses;
- Increase the reporting of incidents through the dissemination of information on available reporting mechanisms;
- Support the development of evidence-based policies and actions to combat online discrimination.

Training and Capacity Building

In September 2024, a four-day *Training for Trainers* session was held in Varna, focusing on the emergence, escalation, and counteraction of online hate speech. Participants from Bulgaria included Ahmed Hasanov, Gamze Maksimova, and Maria Hristova, who later transferred the acquired knowledge to Diyan Dankov, a lawyer and legal advisor at Association Integro with extensive experience in human rights and anti-discrimination work, and to Vergil Ibryam, an activist experienced in national studies on hate speech (2010–2014).

Between 26 and 29 August 2025, the national monitors conducted a national capacity-building training attended by 22 participants. The sessions aimed to:

- Identify the root causes of hate speech;
- Reflect on participants' own biases;
- Understand trends in the dissemination of antigypsy hate speech online;
- Master tools for reporting hateful content;
- Analyze key thematic areas of antigypsy speech related to housing, socio-economic conditions, education, and cultural aspects.

Legislative and National Context

In Bulgaria, combating hate speech and antigypsyism is governed by a range of national laws and institutional mechanisms. The Penal Code criminalizes public incitement to violence or discrimination on ethnic grounds, while the Anti-Discrimination Act explicitly recognizes antigypsyism as a form of racial discrimination. The Personal Data Protection Act and the Electronic Communications Act regulate the processing of personal information in digital environments. The Commission for Protection against Discrimination (CPD) handles complaints and cases related to hate speech, despite limited authority over content moderation on private

platforms. The National Council for Electronic Media (CEM) also monitors and sanctions online media content that incites discrimination or violates human dignity.

At the European level, the Digital Services Act (DSA) and the Directive on Combating Racism and Xenophobia require platforms to take action against discriminatory and hateful content. The AI Act (2024) introduces transparency and accountability requirements for automated content dissemination, prohibiting algorithms that could generate discrimination or profiling based on ethnic or religious grounds. These European mechanisms complement national legislation and provide a legal framework for addressing antigypsy hate speech online.

Significance and Contribution of the Study

The study provides valuable insights into the challenges faced by Roma communities in digital spaces and the mechanisms available to counter online discrimination. The data support evidence-based policy development, public awareness-raising, and capacity-building among activists, legal professionals, educators, and young Roma internet users. Monitoring demonstrates that hate speech often begins with seemingly "harmless" jokes or memes that normalize stereotypes and, if left unchecked, can escalate into discrimination and even violence.

The TAAO project illustrates that effectively combating online antigypsyism requires not only legal regulation but also education, awareness of personal biases, and the cultivation of a digital culture grounded in respect and empathy. By training monitors, facilitating knowledge exchange, and collaborating with institutions, the initiative promotes active citizen engagement and strengthens democratic values in digital environments.

Examples and Trends in Bulgaria (2023–2025)

Media Cases and Public Statements:

- During 2023–2024, public figures made statements with antigypsy undertones, sparking extensive discussion in the media and on social networks. This indicates that hate speech is not solely an online phenomenon but is closely linked to public discourse and the political environment.
- **The national monitoring report on hate speech**[1], produced by Integro Association based on data from the *Peer Education to Counter Antigypsyist Hate Speech Online* (PECAO) project, shows that a significant portion of publications contain stereotypes and generalizations (e.g., "Roma are always poor," "they do not want to integrate," "Roma rely solely on social benefits," "they do not pay taxes and live off the state"), creating a foundation for normalized discrimination.

[1] This report provides a national monitoring analysis of antigypsy hate speech in online environments, developed within the framework of the PECAO project (*Peer Education to Counter Antigypsyist Hate Speech Online*). It combines a peer-education methodology with qualitative monitoring, aiming both to influence the attitudes and behaviors of young Roma individuals and to systematically understand the manifestations of online antigypsyism, identifying key themes, forms of aggressiveness, and patterns of racial discrimination in digital spaces.

- Social media activity has increased, with groups and pages of unclear origin spreading memes that mock Roma people or employ offensive slang, a trend confirmed by Integro's analysis.

Statistical Findings from Association Integro's Report:

- Approximately 60% of the analyzed publications related to Roma communities used negative framing, while only 15% presented the community positively or in a balanced manner.
- Online comments frequently demonstrate double standards, for instance criticizing social benefits for Roma without considering structural factors such as historical discrimination or educational barriers.
- Platforms often respond slowly, or not at all, to content violating media ethics and human rights principles, echoing findings from European-level analyses of hate speech moderation.

Legislative and Institutional Frameworks:

National Level:

- **Penal Code** – criminalizes incitement to violence or hatred based on ethnicity, religion, or other protected characteristics.
- **Anti-Discrimination Act** – defines and provides mechanisms against discrimination, including ethnic discrimination, with complaints handled by the Commission for Protection against Discrimination (CPD).
- **Media and Online Platform Legislation** – provides a regulatory framework for content, although significant legal and technological gaps remain for social media.

European Level:

- **Directive 2000/43/EC (Framework Directive on Racism and Xenophobia)** – requires Member States to criminalize certain forms of discrimination and hate speech.
- **Digital Services Act (DSA)** – imposes obligations on platforms for content moderation, transparency, and complaint mechanisms.
- **AI Act** – introduces requirements for algorithms to prevent discriminatory profiling based on ethnicity or other protected characteristics.

These legal frameworks enable coordinated approaches among states, civil society, and online platforms; however, implementation remains challenging, as evidenced by findings from Integro's report.

Analysis of Countermeasures and Recommendations:

- Moderation and reporting of content remain underdeveloped: many users do not know how to submit complaints, the process is slow, and outcomes are unpredictable.
- Trainings conducted within TAAO are critical for building capacity among activists, legal professionals, and youth to identify "soft" forms of hate (memes, mockery, irony) and their potential escalation.
- **Recommendations:**
    - Strengthen collaboration between the CPD, online platforms, and civil society through long-term partnerships and "trusted flagger" schemes.
    - Adapt training for teachers, mediators, and youth workers to recognize and respond to online antigypsyism.
    - Systematically collect and analyze data on hate speech cases to guide targeted interventions.
    - Increase transparency and accountability of online platforms through public moderation reports and responsible use of automated tools in compliance with the AI Act and DSA.

1.3 Monitoring Methodology

The monitoring process was based on the direct observations of trained monitors and the universal **methodology developed by Prof. Rotaru**[2]. Each case was identified, classified, and analyzed using the structure of the Monitoring Tool (MT) – see Appendix 1. The activity encompassed both quantitative and qualitative content analysis.

The project provided participants with practical experience in combating hate speech and a sense of belonging to a shared cause. Participants came to understand that hate often originates from seemingly "innocent" jokes or posts, whose authors may not recognize the harmful consequences of their words.

1.4 Significance and Objectives of the Report

Early detection and mitigation of online hate speech are crucial for building a safe digital environment.

This report presents a quantitative and qualitative analysis of online content targeting the Roma community in Bulgaria and highlights the efforts of young monitors actively opposing hate speech.

[2] The tool for monitoring and analyzing online hate speech publications was developed based on the proprietary methodology of Assoc. Prof. PhD Habil. Ileana Rotaru. It is part of the TAAO project (*Together Against Antigypsyism Online*), a continuation of the PECAO project (*Peer Education to Counter Antigypsyist Hate Speech Online*), and aims to systematically track, assess, and counter anti-Roma attitudes and online hate speech across social media and other digital platforms.

The report aims not only to provide statistical data but also to demonstrate that change is possible when young people choose to act. Indifference is not an option—every response to hate matters.

2. METHODOLOGY

2.1 Purpose and Design

The study conducted in Bulgaria within the framework of the *Together Against Antigypsyism Online* (TAAO) project was designed based on a combined qualitative and quantitative analysis of antigypsy attitudes and online hate speech. The primary focus was on social media platforms—Facebook, TikTok, X, YouTube, Instagram—as well as smaller but active digital platforms.

The monitoring covered the period from October 2024 to September 2025, applying a universal European methodology for identifying, tracking, and reporting hate speech and antigypsy content. The main data collection tool was a specialized Monitoring Tool (MT) used by all partner organizations.

Main Objectives of the Study:

- **Identify explicit and coded forms of antigypsy hate speech.**
- **Analyze public discourse and negative portrayals of Roma communities in online spaces.**
- **Track the dissemination of negative publications, commenting dynamics, and audience profiles.**
- **Assess the impact of negative comments on the broader public environment.**
- **Observe platform responses following reports of violations.**
- **Monitor the number and types of removed publications as an indicator of moderation effectiveness on social media.**

A total of 528 publications containing explicit or coded antigypsy content were identified. All cases were reported through a dedicated Google Form created within the project and, in parallel, directly to the platforms. The Bulgarian monitoring team systematized the data in a comprehensive spreadsheet, enabling tracking of the proportion of reported content that was removed. This served as an important quantitative indicator of moderator effectiveness, particularly given that many moderation processes rely on algorithms with limited accuracy in assessing hate content.

The collected quantitative and qualitative information was synthesized into four interim quarterly reports, highlighting key observations and trends.

Reports were shared via an online platform accessible to all partners in the six European countries. The data were centrally analyzed by the Hungarian partners, Roma Versitas, who produced international summaries for each reporting period.

Reporting Periods:

- October–December 2024 (uploaded on 30.01.2025)
- January–March 2025 (uploaded on 30.04.2025)
- April–June 2025 (uploaded on 05.07.2025)
- July–September 2025 (uploaded on 28.10.2025)

The experience gained from the previous PECAO project significantly facilitated the process. All partner teams, including Integro, refined their work through regular meetings, team discussions, and expert feedback from Prof. Ileana Rotaru. In some cases, overlap occurred in publications reported by different monitors, but discrepancies were promptly clarified and resolved, contributing to higher data quality and consistency.

2.2 Participants – Bulgarian Monitoring Team

Integro formed a team of five young Roma individuals who are active social media users and possess diverse professional and educational backgrounds. The team included a lawyer, a municipal administration specialist, current high school students, a university student, and a long-standing activist. All members demonstrated strong motivation and commitment to addressing and countering hate speech.

Profiles of Youth Monitors (Bulgaria)

Participant	Background and Education	Previous Experience	Motivation for Participation
Diyan Dankov	36, lawyer, Razgrad; married with two daughters	15 years of experience with Integro; team leader for volunteer and youth programs	Long-standing interest in human rights; desire to contribute professional expertise
Maria Hristova	22, Kotel; Senior Specialist in "Culture and Social Activities"; student of English Philology	Participation in PECAO and volunteer monitoring	Active online; aims to support real change and stronger control over hate content
Gamze Bagryanova Maksimova	19, Razgrad; graduated from PMG; interested in pedagogy	Participation in previous Integro projects	Motivated to counter hate speech and inspire other young Roma

Participant	Background and Education	Previous Experience	Motivation for Participation
Ahmed Hasanov	17, Razgrad; high school student at PMG, "Applied Programming"	Participation in SafeNet project (2022–2024); reporting discrimination cases	Desire for a safer online environment; personal solidarity and responsibility
Vergil Ibryam	36; married; long-standing activist and former Integro coordinator; student at RU	Experience in monitoring media representations and hate speech	Commitment to Roma rights and high-quality monitoring

All monitors participated in a two-week online consultation seminar led by Prof. Rotaru, focused on identifying, categorizing, and reporting hate speech within the context of national legislation and the project's methodology.

In September 2024, part of the team attended a four-day training camp, which further strengthened their capacity. Later, they shared the knowledge acquired with other young people during a national training seminar held from 26–29 August 2025 in Varna.

Vergil Ibryam, as an experienced member of Integro, took responsibility for the internal coordination of the monitoring team, while accountability was maintained through continuous correspondence and monthly online meetings. Together with the Work Package 4 coordinator, Kamen Makaveev, they led the process, provided regular feedback, and adapted the methodology as needed.

2.3 Tools and Measures Used During Monitoring – Procedure

The study was based on a Monitoring Tool—a structured analytical instrument adapted from the methodology of the PECAO project (2020–2022). It enables national and international monitors to document, analyze, and reflect on publications containing hate speech or antigypsy content.

The Bulgarian team was tasked with reporting a minimum of eight cases per month. All cases were recorded using Google Forms in accordance with the guidelines provided in Appendix 1. At least three of the reported publications were described in detail to ensure a qualitative analysis.

The highest number of cases containing proven antigypsy content was found on Facebook, TikTok, YouTube, and Instagram. In many instances, the hate content was not limited to the original post itself, but largely emerged from the comments beneath it.

Monitors shared a key observation: even when news was reported neutrally, including positive coverage, the presence of visually recognizable Roma individuals triggered waves of negative, stereotypical, and hateful comments. This trend was confirmed by international partners as well,

highlighting deeply entrenched prejudices, often expressed by people with little or no direct contact with Roma communities.

This underscores the seriousness of the situation and the urgent need for national and international measures to limit the spread and impact of online hate speech.

The Monitoring Tool consists of six analytical sections:

1) **Basic Information** – author, recipient, platform, accessibility
2) **Main Topic** – themes and triggers of the publication
3) **Content Type and Style** – textual, visual, emotional tone
4) **Intensity and Level of Hate** – explicit or coded
5) **Response Measures** – reporting, counter-reaction, replies
6) **Personal Observations and Follow-up Actions** – monitor reflections, platform response

After processing the data, the monitoring team collectively selected cases to be included in the quarterly synthesized reports. When news publications appeared on multiple platforms, the most widely disseminated and highly rated version was chosen. All tools were systematically incorporated into the national summary reports, which then formed part of the consolidated European dataset coordinated by Amaro Drom e.V.

3. RESULTS

3.1 Quantitative Component of the Study

3.1.1 Summary

The quantitative component of the study provides essential information on the volume and characteristics of analyzed publications containing antigypsy hate speech. During the monitoring period, the Bulgarian team identified and analyzed a total of 528 cases that met the criteria for antigypsy targeting and encompassed various forms of hate speech. The categorization was carried out in accordance with the monitoring tool developed based on the methodology of Assoc. Prof. PhD habil. Ileana Rotaru.

Alongside the quantitative assessment, monitors documented their observations using specially designed protocols intended to support the objectives of the TAAO project. These observations reflect how the monitoring process was perceived by the monitors themselves and provide additional context regarding the dynamics of the online discourses under study.

The primary aim of this analysis is to present how antigypsy discourse spreads in online environments, describe the most common manifestations of hate content, and outline the potential impact of such publications on social media audiences. The quantitative section of the report is structured around six analytical dimensions of the monitoring tool:

A. General Information
B. Main Topic and Representation
C. Content Type and Style
D. Non-Textual Elements and Visual Materials
E. Intensity and Levels of Hate
F. Counter-Reactions and Emotional Responses

The processing of reports proved particularly indicative. It was observed that initial reports are usually reviewed by automated, AI-based systems. Only in many cases, following a request for reconsideration, does a human moderator intervene to properly evaluate the content. This two-tiered system slows down the removal of posts containing explicit hate speech and creates conditions for wider dissemination, representing a systemic issue, especially given the high volume of unmoderated hateful content.

It is also important to note that a significant number of publications with antigypsy hate speech continue to circulate freely, reaching wide and often vulnerable audiences. The consequences are particularly concerning, as such content exerts psychological pressure on young people, including young Roma, who may experience fear, anxiety, and insecurity. In some cases, negative stereotypes are internalized, potentially leading to self-limiting beliefs or adaptation to distorted societal portrayals of Roma.

These observations clearly demonstrate the systemic nature of the problem and underscore the need for targeted and coordinated measures. Stricter regulation is required at both the national legislative level and through regulatory bodies to ensure effective enforcement of existing mechanisms for protection against discrimination and hate speech. Simultaneously, social media platforms, which are used by millions of people, must take a more active role in preventing the spread of toxic content. Given that the online environment has the power to shape attitudes and create narratives—both positive and negative—the responsibility for minimizing harmful impacts is shared among institutions, legislators, and the platforms themselves.

3.1.2 Quantitative Analyses

A. General Information – Total Reported Cases and Social Media Platforms: Account and Profile Types

A.1 Total Number of Identified Cases and Social Media Platforms

The total number of documented cases throughout the monitoring period amounts to **528 publications**, distributed across five major social media platforms as follows:

- **Facebook:** 332 cases (62.88%)
- **Twitter (X):** 38 cases (7.2%)

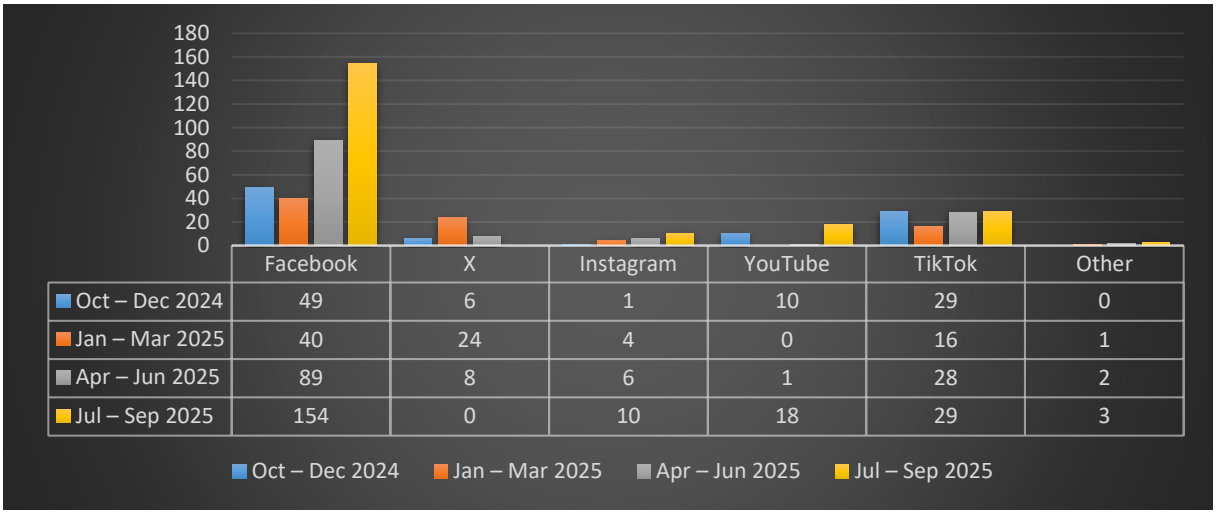
- **Instagram:** 21 cases (3.98%)
- **YouTube:** 29 cases (5.49%)
- **TikTok:** 102 cases (19.32%)
- **Other media:** 6 cases (1.14%)

Table 1. Number of Publications by Quarter

Social media platform	*Oct – Dec 2024*	*Jan – Mar 2025*	*Apr – Jun 2025*	*Jul – Sep 2025*	*Summe*
Facebook	49	40	89	154	**332**
X	6	24	8	0	**38**
Instagram	1	4	6	10	**21**
YouTube	10	0	1	18	**29**
TikTok	29	16	28	29	**102**
Other	0	1	2	3	**6**

Diagram 1 visualizes the distribution of anti-Roma hate speech across social media platforms, clearly showing that the majority of cases originate from Facebook.

Note: During the monitoring process, the Bulgarian monitors collectively agreed not to report cases older than six months. This means that the actual number of online publications containing anti-Roma hate speech circulating on the internet is likely significantly higher than the figures presented here.



The analysis shows that the majority of anti-Roma hate speech cases were disseminated on the social media platform **Facebook (62.88%)**, followed by **TikTok (19.32%)** and **YouTube (5.49%)**.

A.2 Types of Accounts and Profiles Disseminating Anti-Roma Hate Speech

In addition to analyzing the social media platforms through which hate speech is disseminated, the study aims to identify the profiles of users who most frequently publish and share such content. In line with the methodology, the authors of posts are tracked to determine the main distributors of anti-Roma hate speech. Data are recorded directly in the monitoring tool and included in the quantitative analysis. This approach allows for measuring the intensity of posting and mapping the characteristics of profiles that are most actively spreading anti-Roma content online.

Table 2 – Main Profiles Disseminating Online Hate Speech:

Social Media platform		Type of the account/profile				
		Personal/ Private accounts	Online Media * outlets (news, magazines etc.)	Institutional Public accounts	Public figures	Social Media Influencers
a		b	c	d	e	f
1	Facebook	269	24	4	6	9
2	Twitter (X)	36	0	0	0	2
3	Instagram	20	0	0	0	1
4	YouTube	19	5	0	1	4
5	TikTok	83	10	0	0	9
6	Other	0	6	0	0	0

The analysis of the collected data clearly shows that personal accounts of individual users constitute the primary source of anti-Roma hate speech online. Out of the total 528 cases examined,

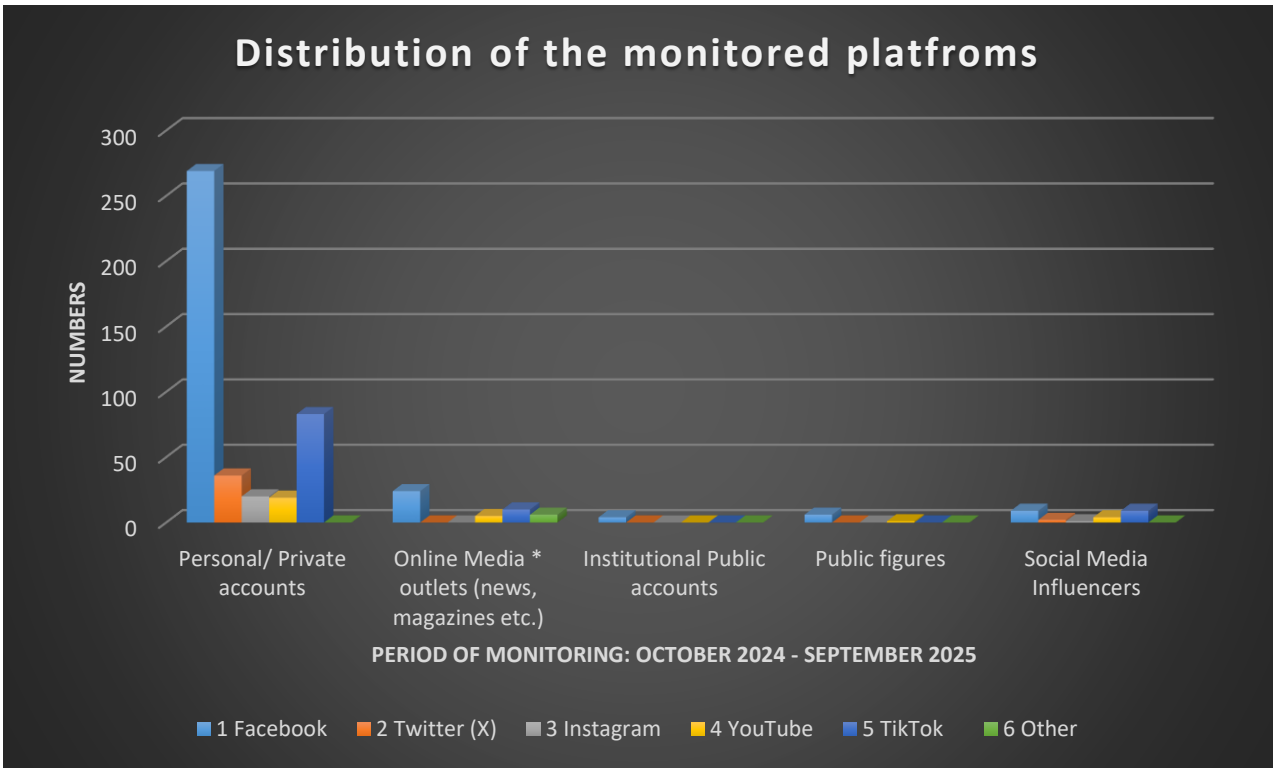
475 posts originated from individual profiles, representing 90.06% of all documented instances. This high proportion highlights that the dissemination of such content is largely driven by mass participation of individual users, rather than organized groups, institutions, or official pages, who actively generate and share hateful narratives.

The second largest source comprises online tabloids and media outlets with official social media profiles. These accounted for 45 posts (8.52% of all cases), primarily distributed through Facebook, YouTube, and TikTok. Although their share is significantly lower compared to individual accounts, media posts have the potential for greater reach and can amplify the visibility and perceived legitimacy of anti-Roma messages.

The following chart (Diagram 2) visualizes the distribution of all registered posts by account type and social media platform. The data clearly underline the dominant role of individual accounts in spreading anti-Roma hate speech online and emphasize the need for targeted measures addressing both individual users and media profiles that contribute to the maintenance and amplification of such narratives.

Diagram 3. Distribution of documented anti-Roma posts by account type and social media platform

The chart illustrates the share of posts originating from individual user accounts, online tabloids, and media outlets across the main social media platforms. Individual accounts dominate the distribution, highlighting their central role in spreading anti-Roma hate speech, while media profiles contribute to amplification and wider visibility of these messages.



B. Main Theme and Representation:

In the context of this report, the analysis of the main themes of anti-Roma posts and the ways in which hate speech is disseminated is of key importance. Previous experience gained during the implementation of the PECAO project highlighted the need to systematize the main thematic areas that shape the hateful narrative against Roma people.

Based on the monitoring methodology used in the TAAO project for tracking anti-Roma attitudes and hate speech in online spaces, we identified ten main thematic categories, classified on a ten-point scale:

1) **Crimes committed by Roma**
2) **Social aspects** – housing conditions, social benefits, poverty, migration, etc.
3) **Educational aspects** – school dropout, learning conditions, scholarships, etc.
4) **Health and sanitary aspects** – pandemics, access to hospitals, abortion, etc.
5) **Social movements and NGOs** – protests, civil rights, representation
6) **Politics** – political representation, parties, elections
7) **Roma leaders, including women**
8) **Cultural events** – music, films, theater, etc.
9) **Sports events** – competitions, games
10) **Other** – additional topics related to Roma, specified when completing the monitoring form (Appendix 1)

These categories provide a clear framework for identifying the most common manifestations of online racism and the formation of stereotypical images of Roma people. The quantitative data presented in this report reflect the frequency of anti-Roma themes and indicate which of them dominate the online space.

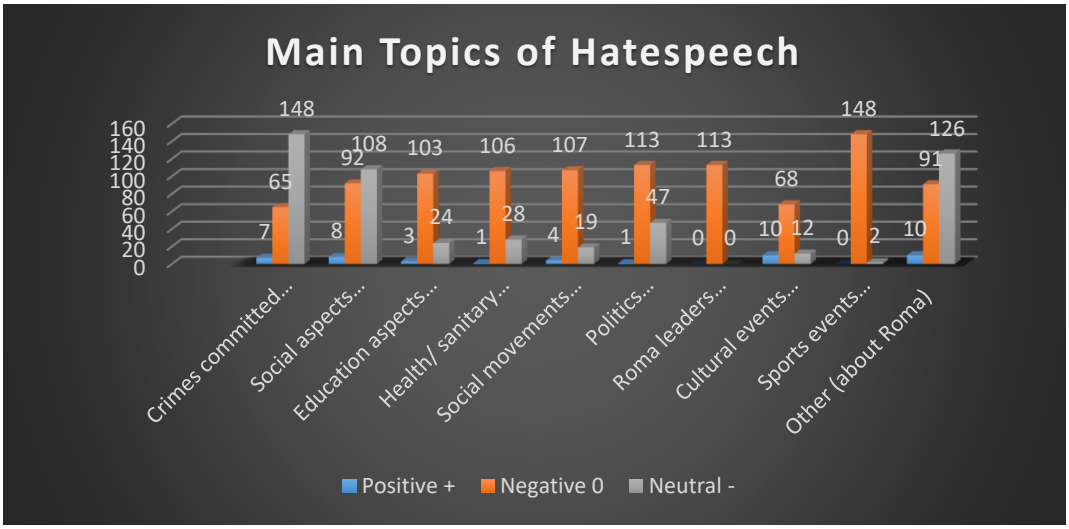
The following table (Table 3) presents the results of the analysis of the main themes in 528 posts featuring Roma. Each post may fall into more than one category, depending on how the author presents the information. For example, a single post may simultaneously address poor housing conditions (social aspect), receipt of scholarships (educational aspect), and political representation of Roma (political aspect).

As a result, the numbers in the table exceed the total number of posts. Nevertheless, they provide a clear quantitative overview of the distribution of positive, neutral, and negative narratives associated with the image of Roma people on social media and online platforms.

Table 3:

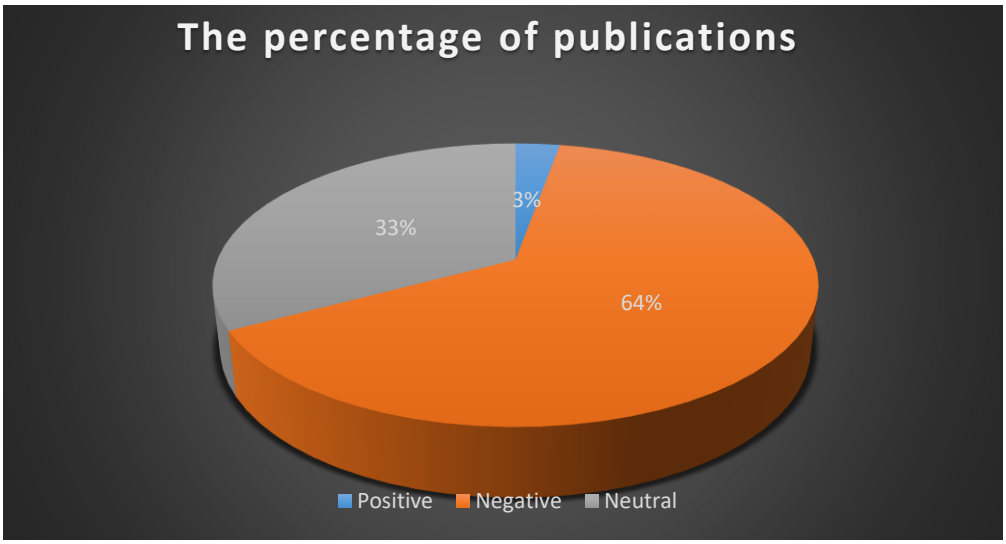
Main topic		Positive +	Negative 0	Neutral -
Crimes committed by Roma	a	7	65	148
Social aspects (housing, welfare, poverty, immigration etc.)	b	8	92	108
Education aspects (drop-out, schooling conditions, scholarships etc.)	c	3	103	24
Health/ sanitary aspects (pandemia, access to hospitals, abortion etc.)	d	1	106	28
Social movements (protests, civil rights, representation) and NGOs	e	4	107	19
Politics (representations, political parties, elections)	f	1	113	47
Roma leaders (including women)	g	0	113	0
Cultural events (music, films, theatre etc.)	h	10	68	12
Sports events (contests, games)	I	0	148	2
Other (about Roma)	j	10	91	126

Diagram 4 illustrates the quantitative distribution of key thematic categories identified in the analysis of posts collected by the Bulgarian monitoring team during the period October 2024 – September 2025. The presented distribution allows for determining the relative contribution of each theme to the total volume of observed media content, as well as outlining the dominant directions in the public discourse. These empirical patterns provide a basis for examining the structural characteristics of the media environment and tracking the dynamics of public attitudes toward the affected groups and phenomena.



Scheme 1 presents the percentage share of posts classified by tone into three analytical categories – positive, negative, and neutral. This typology offers a summarized view of the evaluative frameworks within which the media situate the examined topics. The share of negatively or positively framed posts serves as an indicator of the dominant interpretive models that potentially influence public perception and political discourse. In this sense, the proportion among the three categories goes beyond a descriptive level and provides an analytical perspective for assessing the

risks associated with the spread of stereotypes, as well as the opportunities to change public attitudes through targeted interventions within the framework of the project.



The analytical results outline a complex and multi-layered profile of the online discourse related to the Roma community, while simultaneously revealing structural deficiencies in the current content moderation mechanisms on digital platforms. The presented observations are based on systematic monitoring, where the team's assessments were calibrated according to an established toolkit for analyzing tone and hate speech.

Structural Dominance of Negative Tone

The identified share of publications with a distinctly negative tone (64%) represents a significant indicator of deeply entrenched antigypsyist attitudes in the online environment. These posts are characterized by the use of dehumanizing categories, racial epithets, and various forms of implicit or explicit hostility.
Comment sections under such publications act as "multipliers" of toxicity, containing persistent patterns of extreme verbal aggression, including calls for violence. This requires critical attention in light of both national and European standards for countering racism and hate speech.

Toxicity in the Context of Formally "Neutral" Content

Particularly noteworthy is the fact that even posts with a positive (3%) or neutral (33%) tone trigger pronouncedly negative reactions. This phenomenon reveals a structurally embedded pattern: content that in itself does not violate ethical or professional standards can function as a trigger for mass production of stereotypical and discriminatory comments.
The problem is exacerbated when ethnic origin is mentioned—even when journalistically irrelevant—which automatically activates negative collective perceptions. Empirical data highlight an important regulatory question: to what extent does formal compliance with platform rules suffice to prevent discriminatory impacts on vulnerable groups?

Limitations of Algorithmic Mechanisms and Regulatory Implications

Experience within TAAO clearly shows a significant mismatch between the scale of toxic reactions and the effectiveness of algorithmic systems for detecting and removing problematic content.
Users with private accounts successfully circumvent filters through graphic variations, coded insults, or neologisms that algorithms fail to detect. This points to a structural deficit in platform architecture: moderation systems remain focused primarily on post content without analyzing reactive dynamics in comment sections.
From a regulatory perspective, this raises the need for developing standards that address not only the primary content but also its social consequences and secondary effects.

Impact on Public Discourse and Political Culture

These identified trends have significant consequences for shaping public perceptions. For many online users, the image of the Roma community is constructed almost entirely through negatively toned content or toxic comment dynamics. This supports the reproduction of social prejudices and may facilitate political radicalization, particularly in contexts of social tension.
From the standpoint of European digital democracy policies, these findings underscore the difficulty for existing regulatory instruments to adequately protect vulnerable groups without limiting freedom of expression. At the same time, they emphasize the need for strengthened cooperation among national authorities, platforms, and civil society.

C. Type and Style of Content

The analysis of content type and style provides critical insight into how the image of the Roma community is represented on social platforms and other online channels. This approach allows not only quantitative mapping of posts but also qualitative understanding of the communication practices that shape public attitudes toward the Roma community.

1) **Content Type**
   Posts are classified according to content type and the corresponding tone recorded in the analysis tool—positive, negative, or neutral. The main categories are:

- **Media information** – reports, news articles, analyses.
- **Invitations** – events of cultural, sports, or educational nature, including concerts and webinars.
- **Announcements** – information about daily activities, press releases, and administrative notices.
- **Viewpoints** – editorials, personal opinions, and comments.
- **Advertisements and recommendations** – marketing posts, promotions, job and travel announcements.
- **Curiosities** – unusual or special events, stories, and humanitarian content.
- **Entertainment** – music, videos, films, and other entertainment content.
- **Other** – specified in detail via the media monitoring tool.

This categorization allows identification of dominant online content formats that influence public opinion and the reputation of the Roma community.

2) **Content Style**
Content is also classified by communication style, again distinguished by tone—positive, negative, or neutral:

- **Emotional** – content aiming to elicit an emotional response.
- **Formal/Official** – official statements, professional or administrative texts.
- **Call to action** – posts encouraging specific user actions.
- **Entertaining** – formats intended for amusement, memes, humorous content.
- **Artistic/Fictional/Fantasy** – literary or artistic productions.
- **Scientific** – informational or educational posts based on data and research.
- **Other** – detailed specification via the monitoring tool.

This distinction enables analysis of the communication strategies used online and the potential effect on the formation of stereotypes or positive attitudes.

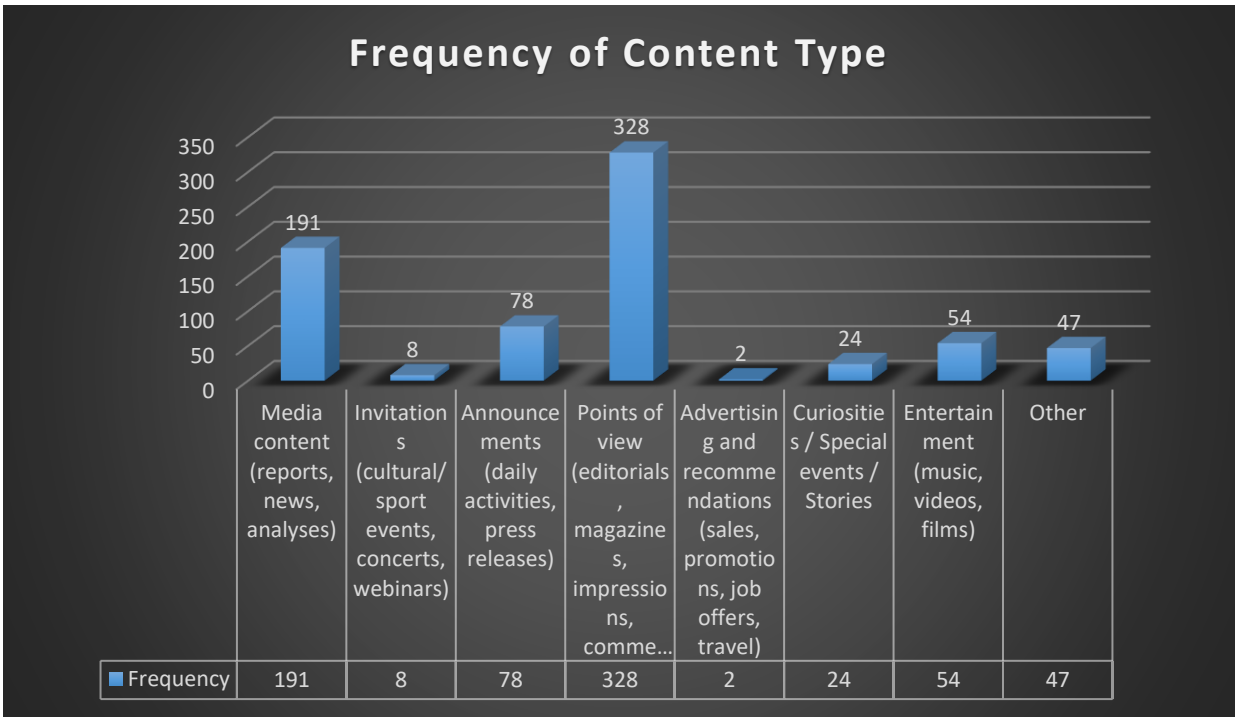
C.1 Analysis of content types

The data presented in Diagram 5 cover the entire monitoring period and provide a systematic overview of the distribution of different content typologies, as well as their contribution to shaping the online discourse, including anti-Roma hate speech.

Note: The numbers shown in the tables and diagrams (e.g., 328, 191, etc.) do not reflect the absolute number of posts, as the total volume of analyzed publications is 528. Each post may belong to more than one content type or style—for example, a news article may simultaneously include the author's personal opinion. Therefore, these values are presented as the "frequency of observed typology," indicating how many times a given typology was identified across the entire set of posts.

This approach allows for a more precise and realistic mapping of the various forms and styles of content, as well as an assessment of their impact on online discourse and the formation of public attitudes.

***Diagram 5* – Frequency of Content Types:**



User-generated opinions are the most frequently occurring content type, with a frequency of 328, representing 44.8% of all documented cases. This indicates that a significant portion of anti-Roma and hate speech content is disseminated through personal posts by users with private social media accounts. Systematic analysis of these posts highlights the role of individual activity in shaping public opinion.

Media publications, including news reports and materials aimed at increasing public awareness, rank second in frequency, with 191 occurrences (26.09%). These are most often posts from journalistic profiles, official accounts of public figures, and influencers, demonstrating the importance of traditional and semi-public channels in shaping online discourse.

Third are announcements intended to inform participants in various social groups (local or thematic), with a frequency of 78 (10.65%). These posts are primarily shared in open or closed Facebook or Instagram groups and target users with specific interests.

Entertainment content, which includes anti-Roma speech in a recreational context, ranks fourth in frequency, with 54 occurrences (7.38%). These posts often evade algorithmic detection due to the complexity of identifying subtle or disguised hate discourse.

Other categories are distributed as follows:

- **Other** – 47 (6.42%)
- **Curious facts** – 24 (3.2%)
- **Event invitations** – 8 (1.09%)
- **Advertising and recommendations** – 2 (0.27%)

Key Conclusions

The distribution of content type frequency provides valuable insight into how different forms and styles of content contribute to the formation of online discourse and persistent stereotypes. The predominance of user opinions underscores the critical role of individual online activity, while the notable share of media publications and announcements highlights the influence of traditional and semi-public channels.

These observations provide guidance for future interventions at multiple levels – platform design, algorithmic moderation, regulatory frameworks, and educational programs – aimed at limiting the spread of toxic content and countering persistent negative stereotypes.

C.2 Graphical analysis of content styles

The graphical analysis of content styles provides a visual representation of the different communication approaches used in online posts related to the Roma community. This type of analysis complements the content-type categorization by highlighting the way messages are framed and the potential emotional or cognitive impact on audiences.

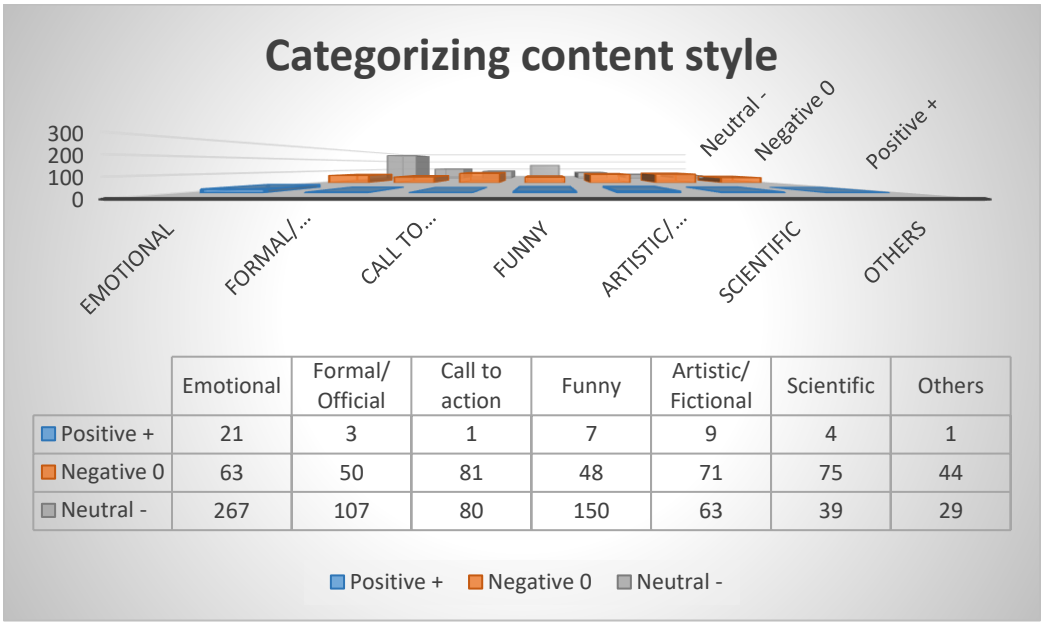
Key style categories analyzed include:

1) **Emotional** – Content designed to evoke strong feelings or emotional engagement from the audience.
2) **Formal/Official** – Posts that convey official statements, professional announcements, or administrative information.
3) **Call to Action** – Content that encourages users to take specific actions, such as participating in events, sharing content, or signing petitions.
4) **Entertaining** – Memes, videos, humorous or playful content, often including subtle forms of anti-Roma speech.
5) **Artistic/Fictional/Fantastical** – Creative outputs, including literary, visual, or fictional portrayals of Roma individuals or culture.
6) **Scientific/Informational** – Data-driven or research-based content, educational posts, or factual reporting.
7) **Other** – Any remaining content styles identified in the monitoring tool that do not fit the main categories.

The graphical representation (Diagram 6) shows the relative frequency of each style, providing insights into which communication strategies are most prevalent in propagating anti-Roma discourse online. Patterns in the graph can help identify which styles are more likely to generate engagement, amplify stereotypes, or influence public perception.

This visual analysis allows for a nuanced understanding of the interplay between content style and audience reception, supporting targeted interventions in moderation, platform policy, and digital literacy programs.

Diagram 6:



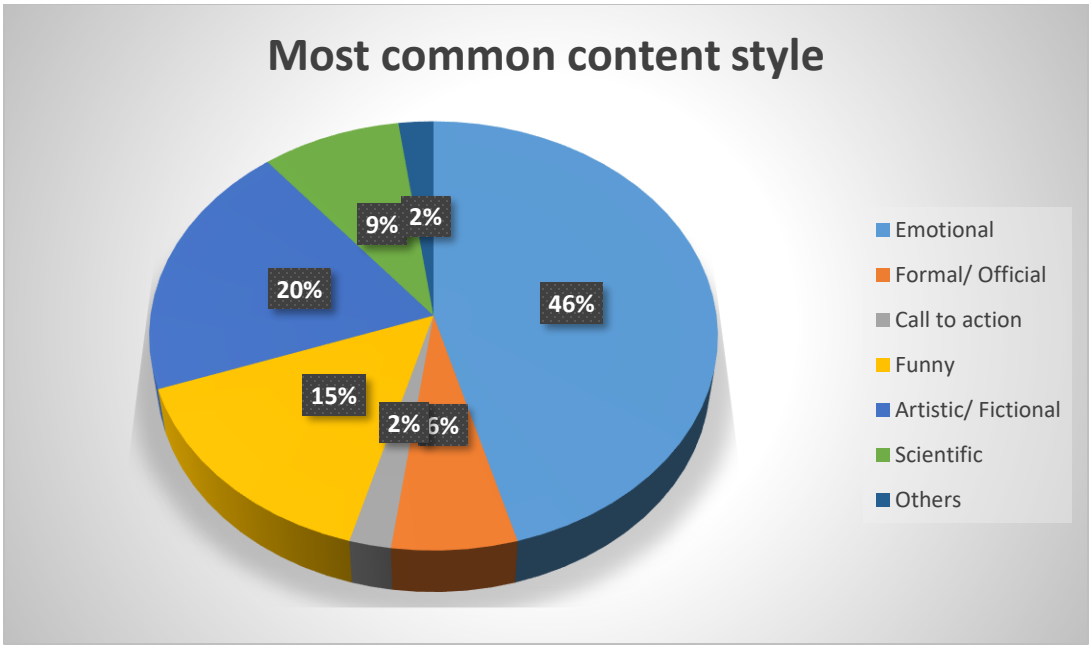
	Emotional	Formal/ Official	Call to action	Funny	Artistic/ Fictional	Scientific	Others
Positive +	21	3	1	7	9	4	1
Negative 0	63	50	81	48	71	75	44
Neutral -	267	107	80	150	63	39	29

Scheme 2 illustrates the distribution of anti-Roma content styles in the online environment.

The most prominent style is **emotional**, accounting for **46% of all cases**. This type of expression is primarily aimed at directly insulting and dehumanizing Roma individuals, while simultaneously mobilizing the support of non-Roma users. These users often reinforce the negative messages through reactions, comments, or by sharing visual elements, which further entrenches prejudices and stereotypes.

The analysis highlights that emotional content plays a key role in amplifying anti-Roma discourse online, demonstrating how affective framing can drive engagement and reinforce discriminatory narratives.

Scheme 2:



The second most common style is the **"artistic"** mode of attack (**20%**), in which Roma artists, public figures, or cultural events associated with the Roma community become the target of negative posts. This style employs creative expressive means, but with the aim of **discrediting and undermining the public image of Roma individuals**.

The third most prevalent is the **"humorous"** style (**15%**), manifested through mockery, ironic memes, illustrations, and humorous texts. Although presented as "jokes," such content plays a significant role in **normalizing hate** and making it socially acceptable.

These diverse styles of anti-Roma expression in the online space reflect the **multilayered nature of online hate**. The predominance of the emotional style indicates that the stereotyped image of Roma is internalized on a personal level by the authors of such posts, who attribute **collective guilt to the entire community without evidence**. This dynamic poses serious risks, as online hate spreads rapidly, is difficult to regulate, and contributes to an environment in which **all Roma are treated as a homogeneous group**, regardless of their individual behavior or actual circumstances.

D. Non-textual elements and visual materials

The analysis of online content with anti-Roma orientation requires systematic monitoring not only of textual posts but also of **non-textual formats**, including images, videos, memes, and cartoons. These visual materials often circulate **without accompanying text**, making them harder for online platforms to automatically recognize and moderate. Many social networks still lack sufficiently advanced algorithms capable of identifying discriminatory content in such formats, creating conditions for the **widespread dissemination of visually coded hate**.

A particularly important observation during monitoring is that the removal of such posts often **requires entirely human intervention**. In numerous cases, initial automated reviews conclude

that the content "does not violate platform standards," necessitating repeated reporting and manual moderator action. This process highlights the **limitations of artificial intelligence** in identifying visual forms of discrimination and the need for **combined mechanisms that include human oversight**.

Monitors also noted a significant number of cases where the **visual content did not match the textual description**. Such posts frequently originate from pseudo-media outlets, populist pages, or profiles of far-right para-organizations that deliberately use misleading images to **create negative connotations about the Roma community**. Frequently, photographs or videos are taken from external sources, media archives, or unrelated events, constructing a **distorted and manipulative interpretation of reality**.

In other instances, even **positive content**, highlighting initiatives or achievements of Roma individuals, triggers a wave of negative reactions, provoked solely by the visual depiction of Roma neighborhoods or Roma faces. This necessitates reporting these posts due to **offensive and aggressive comments beneath them**, which often escalate into hate speech. Such reactions indicate **deeply ingrained stereotyping**, activated even by visual markers that are not inherently negative.

For these reasons, the monitoring methodology also includes the **categorization of visual materials** according to their effect on the audience—positive, neutral, or negative.

Key quantitative findings:

During the twelve-month monitoring period, **528 posts with explicit or latent anti-Roma orientation** were analyzed. Of these, **409 contained a visual element** (77.46% of all posts).

According to the team's observations, posts with visual content **attract significantly more audience attention**, including from Roma individuals, who often react strongly. This indicates a process of **internalized stereotyping**, in which negative visual cues are internalized and reproduced by members of the community.

***Table 6* – Number of Visual Elements and Their Impact Category:**

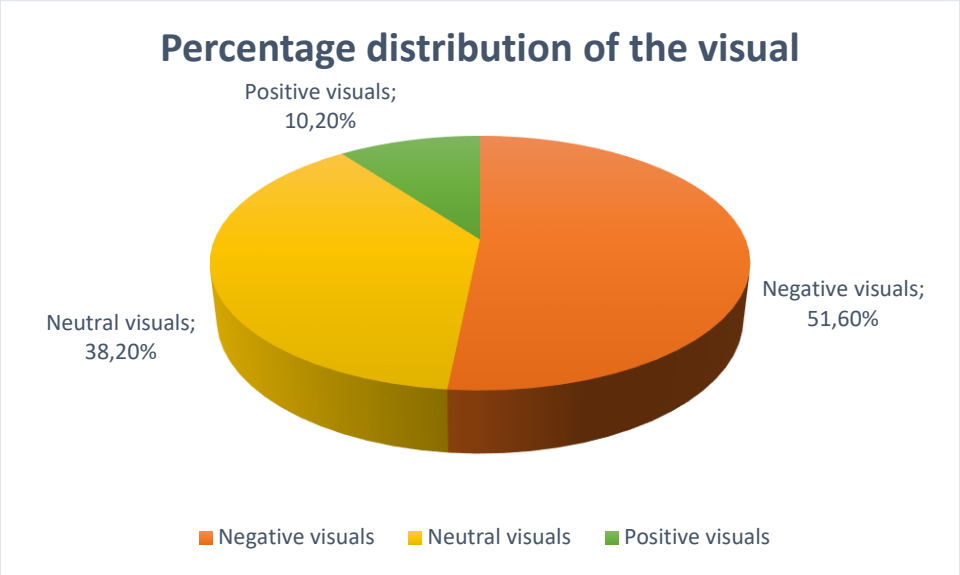
Non-textual form	Positive +	Negative 0	Neutral -
Photos	**8**	**100**	**52**
Memes/Gifts	**1**	**11**	**12**
Caricatures	**0**	**0**	**0**
Multimedia materials (reels, stories)	**4**	**16**	**8**
Videos	**29**	**74**	**81**
Animations	**0**	**10**	**3**

The table shows the distribution of visual elements by tone:

- **Negative visual elements** – 211 items (51.6%), representing the largest group.

- **Neutral visual elements** – 156 items (38.2%).
- **Positive visual elements** – 42 items (10.2%), the smallest category.

This clearly demonstrates that negative images dominate over the other categories.



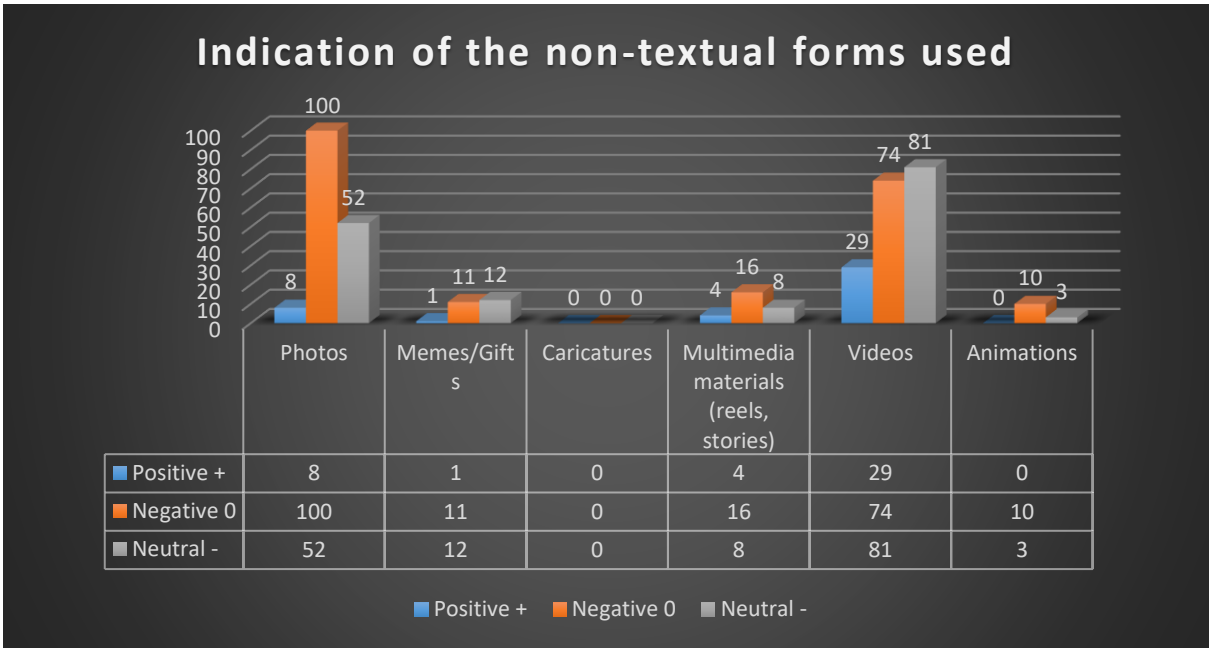
According to the monitors' observations, **even neutral or positive visual materials often trigger negative reactions when they contain images associated with the Roma community**. This demonstrates that visual stereotypes are persistent and can activate hostility regardless of the actual content or message of the post.

Negative reactions and comments under such posts further reinforce harmful stereotypes, creating conditions for the transfer of online hate into offline contexts. This process has already manifested in increasing cases of hate crimes against Roma in recent years, which is a concerning indicator of the social impact of online discourse.

The presented quantitative data and monitoring observations clearly show that visual materials play a key role in the dissemination of anti-Roma discourse. Although they often go unnoticed by moderation algorithms, they have a strong capacity to influence public attitudes, reinforce negative stereotypes, and provoke aggressive behavior.

The following chart illustrates the distribution of types of visual elements that most strongly contribute to the propagation of anti-Roma discourse in the online space.

Diagram 7:



Non-textual and visual materials prove to be a key driver of anti-Roma discourse online. They not only circulate more easily and remain harder to detect automatically, but also provoke strong audience reactions. Monitoring data show that negative visual cues dominate, and even neutral or positive images associated with Roma often trigger hostile comments. This reveals deeply rooted visual stereotyping, which reinforces negative attitudes and contributes to the escalation of online hate.

This dynamic highlights an important transition to the next element of the analysis – the intensity and level of hate generated and fueled precisely through such visual and textual triggers.

E. Intensity and level of hate - ecology and degree of toxicity

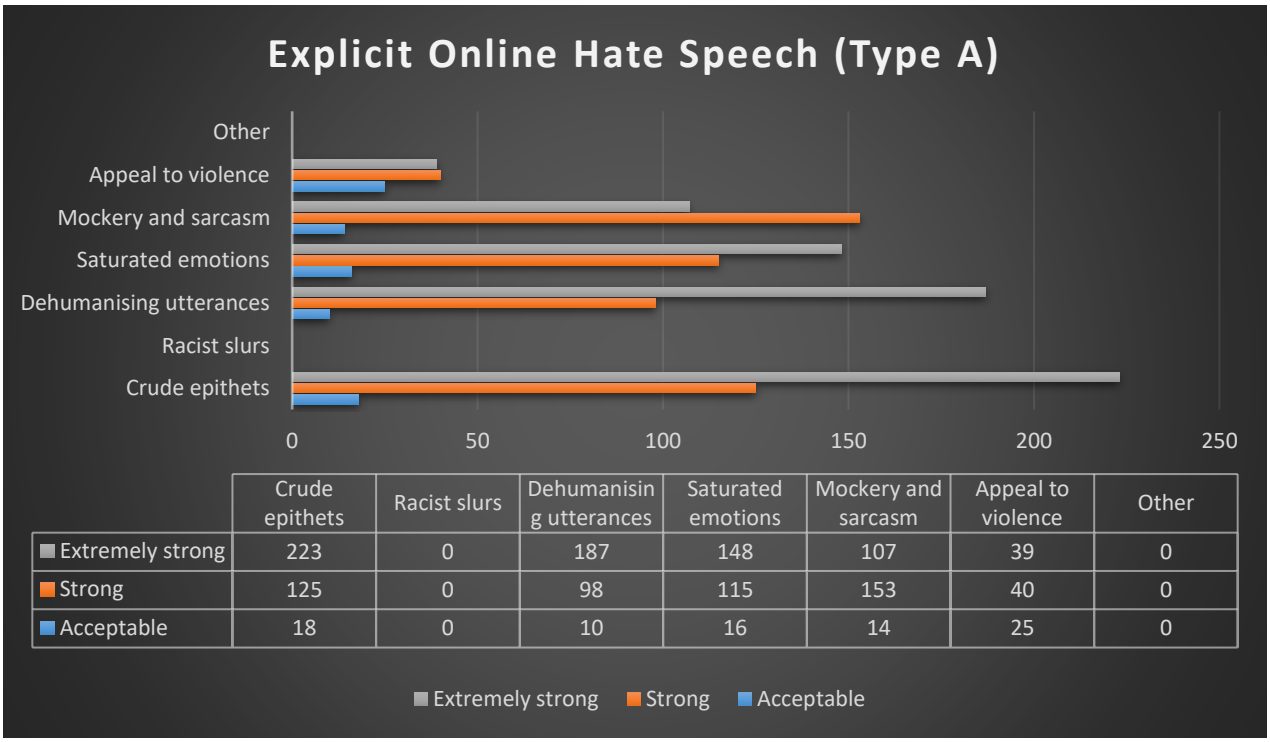
This section focuses on the degree of hostility and the rhetorical forms through which anti-Roma hate speech manifests in the observed online spaces. The study methodology provides for systematic monitoring of two main types of online hate: explicit and latent, with the monitoring tool distinguishing between them by assessing the level of aggressiveness and the linguistic strategies used to conceal or amplify discrimination.

- **Explicit hate (type A)** – expressed through direct and easily recognizable linguistic means, including:
    - ***Crude epithets***
    - ***Racist slurs***
    - ***Dehumanizing utterances***
    - ***Saturated emotions (anger, outrage, hostility)***

- ***Mockery and sarcasm***
- ***Appeal to violence***

This type of content is immediately identifiable and often aims at openly discrediting the Roma community.

Diagram 8: Explicit Online Hate Speech (Type A)

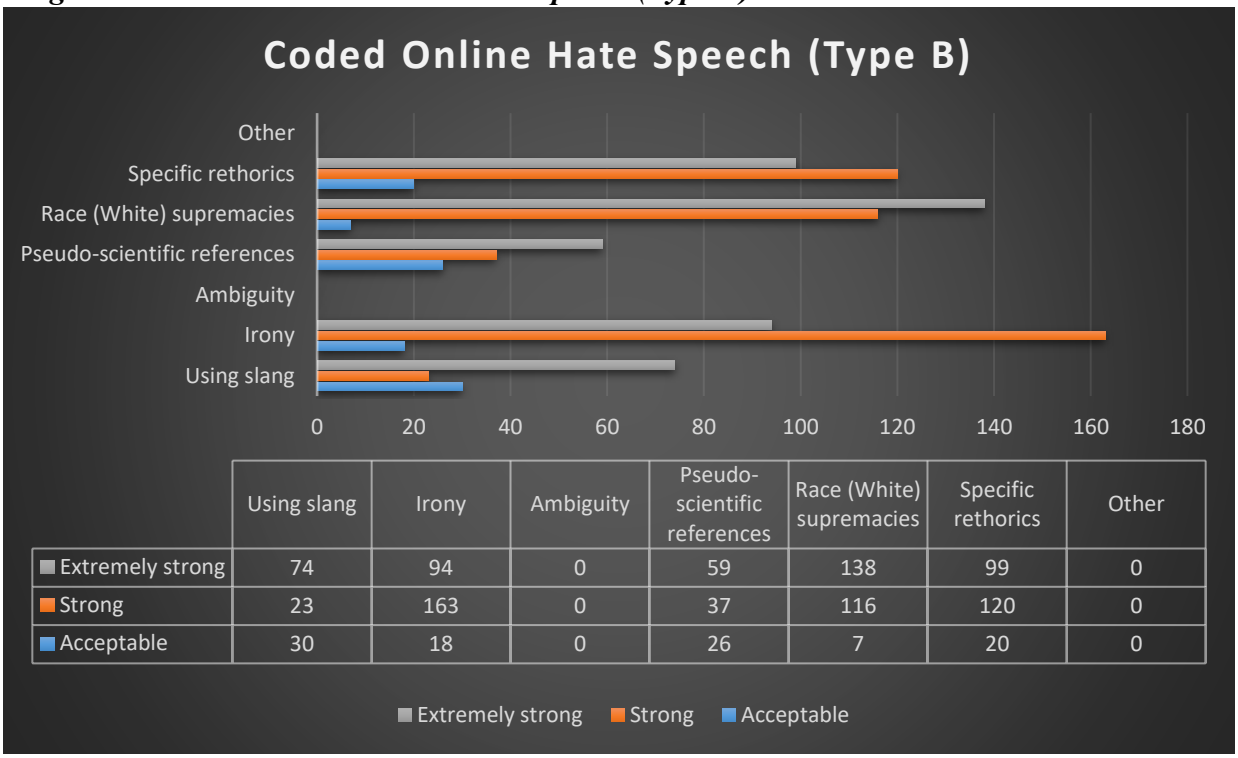


	Crude epithets	Racist slurs	Dehumanising utterances	Saturated emotions	Mockery and sarcasm	Appeal to violence	Other
Extremely strong	223	0	187	148	107	39	0
Strong	125	0	98	115	153	40	0
Acceptable	18	0	10	16	14	25	0

- **Latent hate (type B)** – coded through specific metonymic constructions and expressions, including:
  - ***Use of slang***
  - ***Irony***
  - ***Ambiguity***
  - ***Pseudo-scientific references***
  - ***Race (White) supremacies***
  - ***Specific rhetorical devices***
  - ***Other***

This type of hate is harder to detect and often bypasses automated moderation systems.

Diagram 9: Coded / Latent Online Hate Speech (Type B)



	Using slang	Irony	Ambiguity	Pseudo-scientific references	Race (White) supremacies	Specific rethorics	Other
Extremely strong	74	94	0	59	138	99	0
Strong	23	163	0	37	116	120	0
Acceptable	30	18	0	26	7	20	0

Such differentiation allows for a more precise assessment of the intensity and level of hostility, as well as the mechanisms through which online content reinforces stereotypes and stimulates aggressive attitudes toward the Roma community.

Intensity is measured as the number of mentions of the corresponding type of hate by the monitors in identified cases. Additionally, the monitoring tool includes another key characteristic of anti-Roma content – **level of hate**. This level is determined based on the monitors' observations and subjective perception, and it is classified into three categories: acceptable, strong, and extremely strong.

These categories provide a basis for detailed analysis of the observed cases, allowing the identification of publications with the highest intensity and level of hate, as well as the mechanisms through which anti-Roma messages are disseminated and entrenched online. This approach facilitates systematic examination of trends and patterns of hostility, as well as assessment of their potential social impact.

Explicit and Latent Hate – Intensity Analysis

Explicit hate manifests most strongly through crude epithets, dehumanizing utterances, and saturated emotions, with the highest intensity observed in extremely strong posts.

Latent hate is widely disseminated through ironic and pseudo-scientific expressions, as well as indirect suggestions of racial superiority, which complicates automated moderation.

Quantitative analysis shows that for explicit hate (type A), extremely strong posts constitute **53.41%** of all reported cases. Among these, crude epithets are the most frequent (19.91%), followed by dehumanizing expressions (14.18%), emotionally charged expressions such as anger (11.22%), and mockery (8.1%). These data indicate that extremely strong manifestations of explicit hate account for more than half of all recorded cases, with crude epithets having the largest relative share.

For latent or coded hate (type B), the quantitative analysis shows that extremely strong posts make up **45.3%** of all reported cases. The most common manifestations within this category include coded expressions of racial superiority (13.47%), specific rhetorical devices (15.66%), irony and ironic expressions (9.17%), as well as specific slang and Roma-related words used to demean the Roma community (7.22%). These data indicate that although the proportion of extremely strong cases is lower than in explicit hate, they still represent a significant volume with potential social impact and capacity to reinforce stereotypes.

The overall quantitative analysis highlights the need for combined monitoring and moderation strategies that integrate automated tools with human expertise. The results demonstrate that both explicit and latent anti-Roma speech contain a substantial share of extremely strong posts, requiring a precise approach for identification and management in the online environment.

Identified Categories of Toxic Impact in the Online Environment

These categories reflect how hate speech functions as a social and emotional ecosystem, rather than merely a collection of offensive expressions. Based on observations collected by monitors and data from protocols for analyzing and countering hate speech, six main categories of toxic impact were identified, forming the overall logic of anti-Roma online discourse.

The analysis shows that this type of content operates as a structured system of recurring narratives that reproduce and amplify social exclusion through emotional, symbolic, and ideological mechanisms. Observed publications in the digital public sphere demonstrate a persistent pattern of normalizing discriminatory attitudes, portraying Roma in ways that legitimize hostility and maintain a constant environment of prejudice.

During the analysis, monitors recorded the following frequency of toxic effects, as indicated by full agreement in the reporting table:

1) **Reinforcement and Confirmation of Stereotypes – 282 mentions** - This is the most frequently recorded category. There are persistent attempts to reduce Roma to collective negative traits—"criminals," "lazy," "dependent on welfare." Such simplified narratives erase individuality and reproduce stigma across various online contexts, from comments under news articles to video content on social media.
2) **Deepening Polarization – 217 mentions** - This includes posts and discussions that pit communities against each other through the divisive logic of "us versus them." Polarization

is particularly common in local Facebook groups and political discussions, where anti-Roma rhetoric fits into broader populist and nationalist frameworks. These online "echo chambers" amplify negative attitudes and make racism socially acceptable.

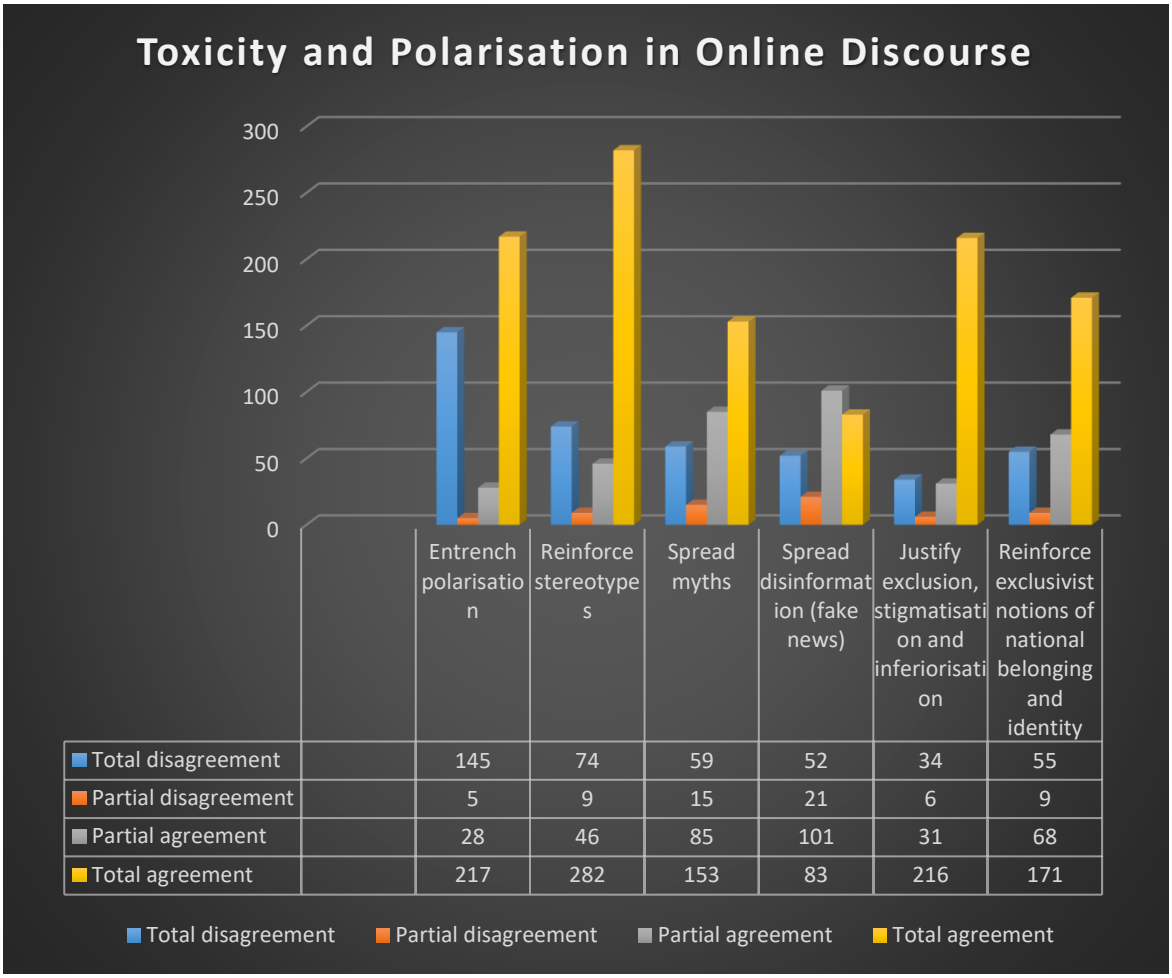
3) **Justification of Exclusion, Stigmatization, and Marginalization – 216 mentions** - Narratives in this category shift the responsibility for discrimination onto the Roma themselves, portraying them as "unable to integrate," "self-blaming," or "incompatible with European values." This denies the structural dimension of inequality and normalizes social exclusion.

4) **Reinforcement of Exclusive Notions of National Belonging and Identity – 171 mentions** - This category includes discourses that associate "Bulgarian" or "European" identity with ethnic homogeneity, and portray Roma as "outsiders" or "threats." This strategy not only sustains anti-Roma sentiment but also broadens intolerance toward other minority and migrant groups.

5) **Propagation of Myths – 153 mentions** - Typical examples include exaggerated or completely fabricated claims about "criminality," "high birth rates," or "privileges" of Roma. These myths perpetuate stigmatization and often go unchallenged by other users.

6) **Dissemination of Disinformation (Fake News) – 83 mentions** - Although the least frequently noted category, it includes significant examples of manipulative content—pseudo-statistical claims, false news articles, or conspiracy insinuations. Their dissemination highlights the audience's vulnerability to unverified information.

Diagram 10 illustrates the dynamics of the ecology of online hate speech targeting the Roma community, visualizing the main directions of this discourse and the most common toxic narratives through which it is stigmatized (**Frozen vs. motile online hate speech**; Lentin, 2016).



F. Counter-reactions and emotional responses

The Bulgarian monitoring team's tool collects data on platform responses and counter-actions undertaken by young monitors after reporting anti-Roma content. This section presents both formal platform measures and the individual strategies employed by young participants when confronted with online hate speech.

As shown by the data analysis (Table 1), the most commonly used response is reporting the post or its author to the platform's system (421 "Definitely Yes"), reflecting high user trust in existing mechanisms and awareness of reporting procedures.

Other frequently applied strategies include:

- **Critique or exposure of the author's tactics/strategy** (87 "Definitely Yes") – used for posts with manipulative rhetoric, stereotypes, or disinformation.

- **Assessment of the effectiveness, success, or quality of the author's arguments and behavior** (155 "Definitely Yes") – an analytical approach aimed at questioning the legitimacy of the narratives.

Less frequently used informal approaches include:

- **Mockery and ironic responses** to the post or author (5 "Definitely Yes") – functioning as a humorous way to cope with hostility.
- **Reciprocal reactions in the same tone** (33 "Definitely Yes") – reflecting a limited willingness for direct confrontation.

Significantly fewer monitors engage in:

- **Serious dialogue or engagement** (258 "Definitely Yes") – although the number seems high, in the context of the total observations this is a moderately preferred approach, as online debates with authors of hate speech are often perceived as emotionally exhausting or ineffective.
- **Reporting to institutions or seeking legal assistance through civil society organizations** (19 and 11 "Definitely Yes") – relatively rare, reflecting limited trust in the effectiveness of institutional responses.

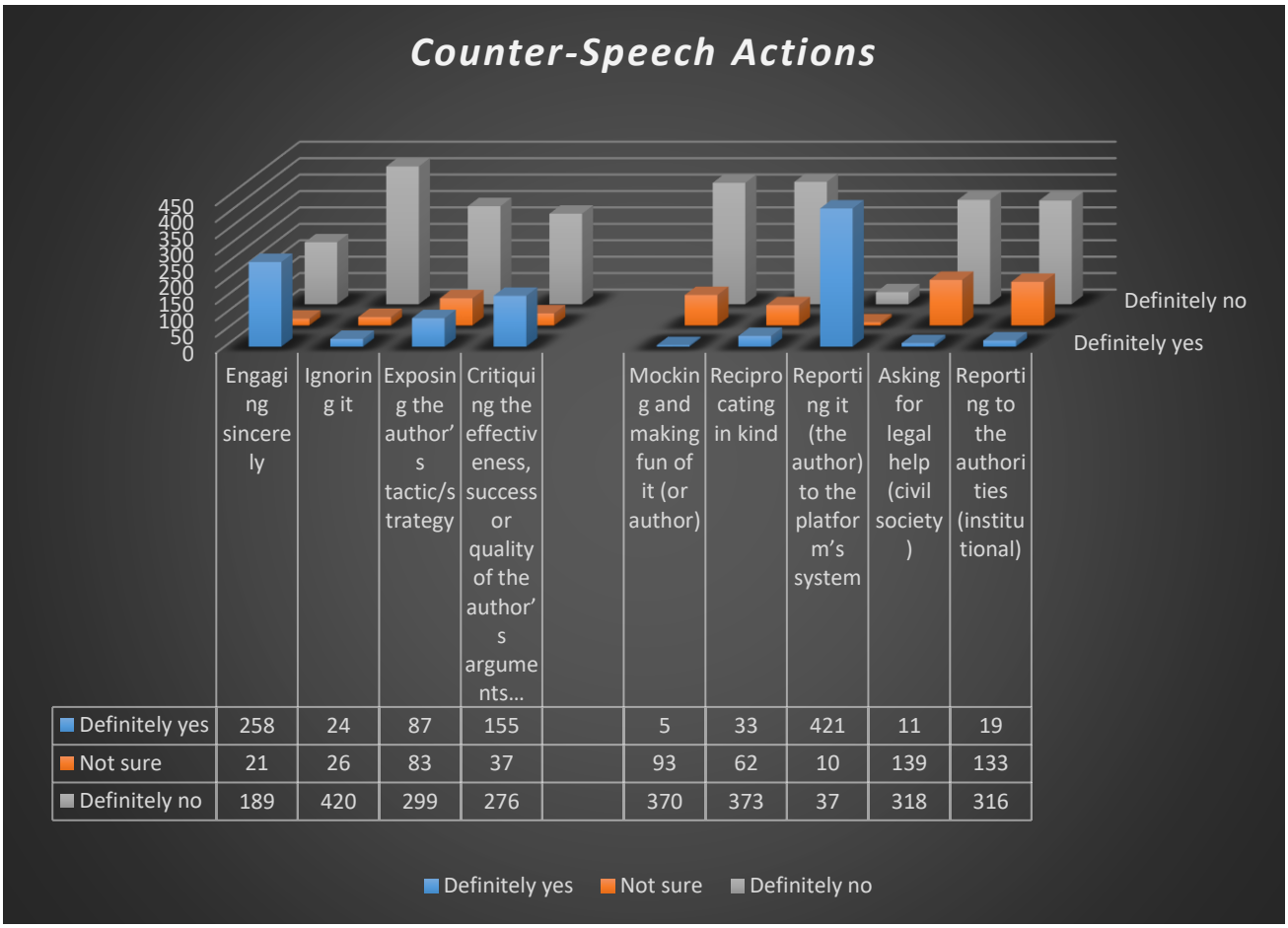
Additionally, a notable portion of monitors expressed uncertainty regarding some strategies ("Not sure"), highlighting uneven application of practices and the need for further training and support.

These results underscore several key observations:

1) **Reporting to the platform** remains the most effective and accessible strategy, providing visible outcomes such as temporary blocking, content removal, or warnings.
2) **Individual and informal approaches**, such as critique, exposure of strategies, or ironic responses, have potential for counter-speech and help develop monitors' analytical and emotional skills.
3) **Lack of consistent feedback from platforms** creates uncertainty regarding the effectiveness of the actions taken, emphasizing the need to enhance transparency and accountability in digital moderation.

In summary, the data from the Bulgarian monitoring team indicate that ecological regulation of online hate requires both institutional mechanisms and active citizen and youth participation, with the synergy between formal reporting and counter-discourse being crucial for reducing toxicity in online environments.

Diagram 11 illustrates the distribution of different strategies used to counteract anti-Roma content online, based on data collected by the Bulgarian monitoring team.



Analysis of Removed Publications and Their Impact on the Data – Supplement to the Quantitative Assessment

Monitoring reported publications required not only a quantitative tally but also careful tracking of whether the content was actually removed or remained accessible in the public digital space. This step serves as a crucial bridge between quantitative and qualitative analysis, allowing for a more precise evaluation of the effectiveness of platform moderation mechanisms.

During the monitoring period, 528 publications containing anti-Roma, toxic, or hate content were identified, all of which breached platform standards. Of these, only 69 publications were removed following user reports, with the majority classified under Typology A: explicit and direct hate speech.

Expressed as a percentage, this indicates that merely 13.06% of reported cases resulted in tangible action by the platforms. This rate is alarmingly low, particularly in light of publicly declared zero-tolerance policies toward hate speech. The limited responsiveness raises concerns about the consistency and efficacy of both algorithmic and human moderation, as well as the platforms' genuine commitment to protecting vulnerable communities.

Given that national monitors reported only content published within the past six months, the actual volume of circulating hate speech is likely substantially higher than the 528 documented cases. In this context, the low removal rate suggests a significant accumulation of problematic content that remains accessible, spreads, and becomes normalized within the digital ecosystem.

This trend constitutes a systemic risk: the online environment is expanding, and the user base is growing, with a considerable portion of young individuals who often lack critical literacy skills to identify manipulative or prejudiced narratives. Consequently, stereotypes and hostile attitudes are internalized as credible representations of the Roma community, despite being frequently detached from direct personal experience.

Overall, these findings not only reveal the intensity of anti-Roma discourse online but also underscore structural deficiencies in digital moderation and highlight the urgent need for targeted interventions—both at the institutional level and through civic, educational, and media-mediated initiatives—to mitigate the persistence and impact of online hate.

3.2 Qualitative component of the study

The qualitative component complements the quantitative findings presented in the previous sections by emphasizing the contextual, emotional, and interpretive dimensions of anti-Roma hate speech, as observed by the Bulgarian monitoring team. Beyond the numerical data, the monitors' accounts and observations reveal the complexity of circulating online hate content, its emotional impact on participants, and the perceived limitations of existing reporting and response mechanisms.

The following sections present the key findings from the qualitative analysis, illustrating the most prevalent discourses, patterns of interaction, and emotional responses among monitors and audiences.

3.2.1 Core Thematic Patterns

Analysis of the 528 documented cases within the Bulgarian monitoring framework outlines a complex landscape of mutually reinforcing anti-Roma narratives. These narratives rarely exist in isolation; instead, they intertwine to create an environment in which hate is reproduced, normalized, and justified.

● Associating Roma identity with criminality, "cultural deficit," and lack of education

The most persistent and frequently observed motif in the monitored publications was the reductionist portrayal of Roma identity as inherently linked to crime, aggression, or social irresponsibility. In a significant number of cases, generalizations such as "they naturally steal," "they don't want to get educated," or "that's just how they are" were used. These claims, often framed as "realism" or "common sense," not only substituted individual responsibility with ethnic identity but also reinforced the notion that social exclusion is justified or inevitable.

● Emotional polarization and construction of moral hierarchies

The observed discourses frequently constructed an explicit "us versus them" opposition, positioning Roma as morally, culturally, and socially "inferior." This moralistic framing served to legitimize aggressive or punitive attitudes, presented as a defense of "public order," "patriotism," or "normality." Such polarization intensified perceptions of threat and contributed to heightened emotional responses, including anger, fear, and disgust.

● Normalization of anti-Roma sentiment through humor, irony, and "everyday jokes"

Another widely observed pattern involved disguising hostile content as a "joke," "funny comment," or "self-irony." This seemingly light and playful tone often masked deep-rooted stereotypes. The use of humor, particularly in the form of memes, allowed some users to spread hate without accountability, as it could easily be dismissed as "not serious," "innocent," or an overreaction on the part of others. The result, however, remained the same: reinforcement of social distance and normalization of prejudice.

● Legitimizing hate through appeals to freedom of expression

In a significant number of cases, hate was justified by invoking "personal opinion," "satire," "right to criticize," or "freedom of speech." This strategy enabled authors to present their anti-Roma claims as part of legitimate public debate rather than harmful content. Monitors observed deliberate exploitation of ambiguities in platform rules and national regulatory frameworks, allowing hostile messages to remain online and circulate without substantial consequences.

The interweaving of these thematic patterns creates an environment in which anti-Roma messages are not merely circulated but actively reproduced and reinforced. "Humorous" forms of hate often serve as entry points, normalizing negative stereotypes and reducing public sensitivity to more overtly hostile messages. This normalization paves the way for harder narratives, such as criminalizing representations of Roma or the moral hierarchization of "us versus them." Subsequent appeals to freedom of expression function as a protective mechanism, allowing all preceding forms to remain in the public sphere without authors being held accountable or even recognizing the harm they cause. In this way, the four patterns mutually reinforce each other, contributing to a resilient and hard-to-penetrate ecosystem of online anti-Roma hate—an observation fully consistent with the findings of the Bulgarian monitoring team under the TAAO project.

3.2.2 Analysis of Selected Cases

Within the monitoring framework, numerous instances of anti-Roma hate speech were identified. For the purposes of this report, we focus solely on highly illustrative and indicative examples. The selected cases encompass both Type A—overt hate speech—and Type B—latent, indirect, or coded aggressive discourse. It should be emphasized that the examples presented here represent only a minimal fraction of all monitored and reported cases.

The following sections present key excerpts and findings from the Bulgarian national report, illustrating characteristic forms of anti-Roma hate and the responses of the digital environment:

Case 1:

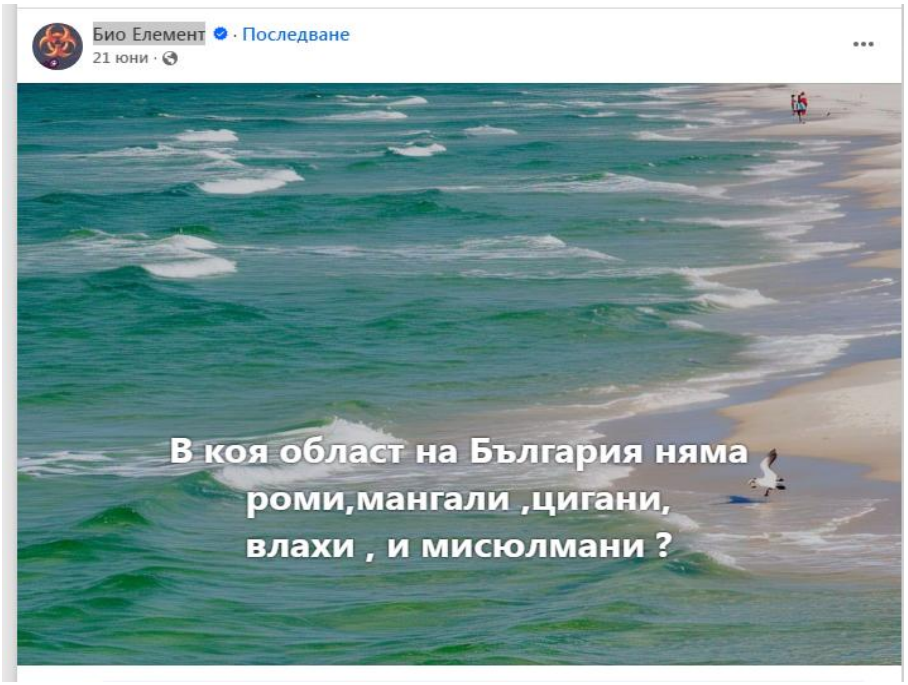
The author of the post is a well-known former Bulgarian journalist and current Member of the European Parliament (MEP), notorious for his inappropriate remarks and repeated public scandals. The case in question is yet another example of his problematic rhetoric. He positions himself as being against the Eurozone, against the euro, and against European values—despite serving as an MEP. He is also openly hostile toward the Roma community.

In a recent public Facebook post, he wrote: *"At the next elections… the fate of Bulgaria will be thrown into one of two paths—Gypsification, corruption…"*

This statement is not only politically inappropriate but also contains explicit hate speech. The term *"Gypsification"* is a **racist and xenophobic expression** used to imply cultural and societal decay allegedly caused by the presence of Roma people. The post promotes harmful stereotypes and fuels division, and its use by an elected political representative makes it even more concerning.

This example represents an overt form of hate speech. While it does not contain visual elements, its impact is highly negative. Despite being reported, the post was not removed from the social media platform, with the justification that it "does not violate the platform's standards." Such content is often treated as "personal opinion," creating a broad gray area that limits the possibilities for effective moderation. An additional concern is that the message was disseminated by a public figure and representative of Bulgaria in the European Parliament—an aspect that amplifies the legitimization and impact of negative stereotypes against the Roma community. Nevertheless, the post continues to circulate online and elicits supportive reactions, highlighting the resilience and normalization of such attitudes within the digital environment.

Case 2:



Всички коментари

Мариана Генчева
Цялата ни държава е цигания, а тая зган скоро ще са повече от българите в България

4w Харесване Отговор 3

Автор
Био Елемент
Мариана Генчева верно е ....и ще дойдат мангали от всякаде заради еврото ...ще стане кат Париж един бял 33 ма черни у Берлин 52 черни 3 ма бели общо взето да пускат атомната бомба викам аз ....и да се свършва чалми евреи афроамериканци 😹😹😹😹🤦‍♀️🤫😁😉😵

4w Харесване Отговор

Дон Чичо
Био Елемент Не забравяй и крепостните пияници от РОССИЯ !

4w Харесване Отговор

The author of the post is most likely the administrator of a public group called “Bio Element” — an influencer page that mainly shares videos and photos of sports cars.
In the post, the author addresses the audience with the question: **“Which region in Bulgaria has no Roma, gypsies, ‘mangali’*, Vlachs, or Muslims?”**

This is not just a casual inquiry. The author is attempting to provoke amusement or entertainment among readers, but by using offensive slurs such as “mangali” and “gypsies,” he contributes to further polarization in society.

It is well known that these terms are derogatory and have no positive connotation for the Roma community.

One of the comments under the post states: ***“Our whole country is one big gypsy mess, and soon this scum will outnumber the Bulgarians in Bulgaria.”***

In response to this comment, the author of the original post replied with the following: ***"It's true… and ‘mangali’ will come from everywhere because of the euro… it will become like Paris — one white person, thirty-three Black people; in Berlin, fifty-two Black people, three white people. Overall, I say they should just drop the atomic bomb and be done with it — turbans, Jews, African Americans."***

This reply contains extremely hateful, racist, and violent language, including:

- The use of the racial slur **“mangali”**, a derogatory term targeting the Roma community.
- **Racist stereotypes** and generalizations about Black people in European cities.
- A **direct call for mass violence**, suggesting the use of an atomic bomb.
- Expressions of **hatred and dehumanization** directed toward **Muslims (“turbans”)**, **Jews**, and **African Americans**.

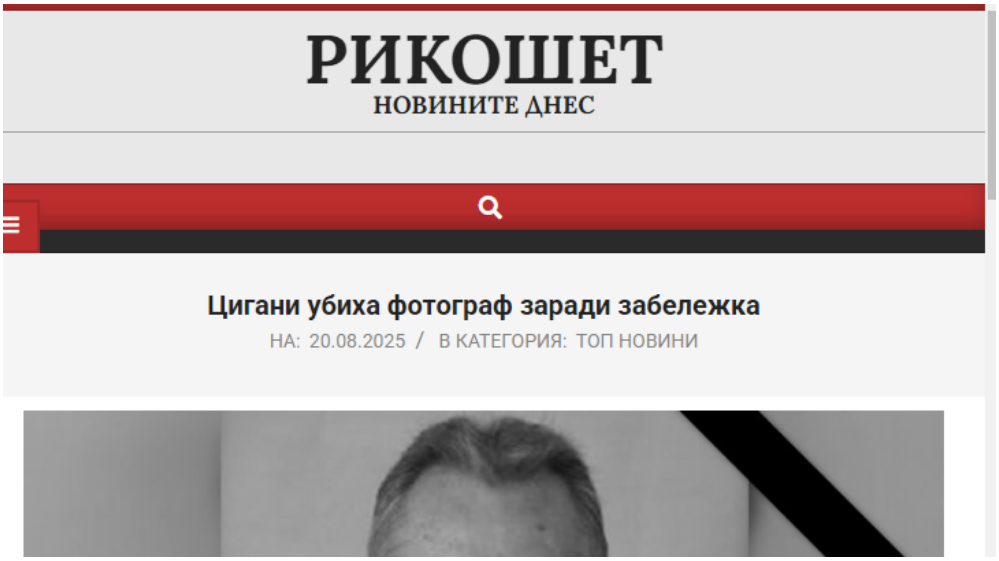
This kind of language constitutes a severe violation of community standards and is a clear example of **hate speech and incitement to violence**. It should be removed immediately, and appropriate action should be taken against the author.

This is clear evidence that such posts should not be allowed on social networks, as they fuel hatred toward Roma people and are a direct example of hate speech.

**** Note:*** *“Mangali” is a deeply derogatory racial slur used against Roma people in Bulgarian.*

Case 3:

Official website of the tabloid Rikoshet, "Top News" section



The news website reports on an incident in which a well-known local photographer lost his life after being beaten by his neighbors, following a complaint he had made about loud music. The content of the article adheres to journalistic ethical standards, with the exception of its headline: *"Gypsies killed a photographer over a remark."* The reference to ethnicity fuels interethnic tension.

This was an isolated case, yet it is presented as representative of the entire Roma community through the phrasing *"Gypsies killed…"*. In the accompanying Facebook post, the text reads: *"He (the photographer) passed away after being brutally beaten by a gang of Gypsies."*

Such wording alone is enough to capture readers' attention and predispose them negatively, even without reading the full article. This is a highly impactful publication with a clear anti-Roma message.

Facebook page

Case 4:

In a recent episode, a well-known YouTuber and influencer, who explores various topics and presents "facts," explains the so-called **true nature of the Roma people**, relying on false evidence, stereotypes, and prejudices that reinforce the negative image of the Roma community not only in Bulgaria but worldwide. The video begins with a **mocking imitation of a Roma person**, who is shown dancing and telling the filmer not to record him. The entire video is filled with stereotypes that generalize the entire Roma community, portraying traits such as **theft, pickpocketing, unwillingness to study, and poor hygiene as their "best qualities,"** thereby reinforcing negative prejudices about Roma people.

Overall, the video is offensive and violates the platform's basic standards, yet it has not been removed from YouTube **because it generates a high number of views** and contains advertising placements, which contribute to the platform's financial incentive not to take it down. **This is a typical example of sarcasm and ridicule – online hate content presented in a humorous and mocking manner.**

3.2.3 Emotional Effects, Online Dynamics, and Key Observations from the Monitoring of Anti-Roma Hate in Bulgaria

During the TAAO monitoring process, the Bulgarian team was systematically exposed to a high volume of anti-Roma content, most frequently on Facebook and TikTok. Hate speech in these

platforms circulated through mockery, negative stereotypes, memes, and short videos framed as "humor" or "everyday observations." This type of content proved particularly challenging to interpret, as hostile messages were often concealed under the guise of a joke, facilitating their normalization and high engagement within the platforms.

Despite prior training, repeated exposure to such materials caused emotional strain among the monitors. Some reported feelings of helplessness and frustration, largely due to the fact that content was rarely removed even after multiple reports to the platforms. This reinforced the perception that algorithms and moderation mechanisms tolerate or underestimate anti-Roma discourse, particularly when it is presented in an "entertaining" format.

During the monitoring, the team identified persistent online ecosystems - pages, groups, and profiles - that systematically reproduce anti-Roma narratives in the form of pseudo-news, humorous collages, videos with "viral" potential, or posts framed as "civic positions." The concentration of such content on Facebook and TikTok is further amplified by platform algorithms that reward popular and provocative materials.

Although the monitors discussed potential reactions, the group did not undertake organized counter-speech or coordinated responses. The primary focus remained on observation and reporting, highlighting that platform self-regulation alone is insufficient to curb these practices.

The collected observations underscore the persistent presence of anti-Roma hate in the Bulgarian online space and emphasize the need for more effective moderation mechanisms, clearer legal definitions of online hate speech, and more systematic support for young people engaged in monitoring and combating discrimination.

4. DISCUSSION

The national study conducted under the TAAO project in Bulgaria reveals a persistent and deeply rooted pattern of anti-Roma hostility in the digital environment, as well as significant discrepancies between the existing legal framework, platform policies, and the actual online experiences of users. The quantitative data make it possible to outline the most common forms and patterns of hostile content, while the qualitative observations deepen the understanding of its psychological, social, and algorithmic mechanisms.

Alongside the technical limitations of the platforms, the analysis also highlights an important human and educational dimension: prejudices, stereotypes, and a lack of empathy that are amplified and normalized online.

4.1 Interpretation of the Results

The monitoring confirms that anti-Roma narratives are widespread across the Bulgarian online space, particularly on Facebook and TikTok, where they easily adapt to current trends and platform algorithms. In addition to overt hate speech (insults, dehumanization, calls for violence), a significant portion of the content appears in the form of "coded," ironic, or pseudo-rational language. These formats evade both legal classification and automated filters, often disguising themselves as "humor," "personal opinion," or "social commentary."

Platform algorithms further amplify the visibility of such materials by prioritizing content that triggers strong reactions. In this way, hate becomes both a product and a tool for generating engagement — a process that gradually erodes public trust and social cohesion.

Particularly alarming is the low share of removed posts after reporting — below 10%. This raises serious concerns about the effectiveness and transparency of moderation systems, which are dominated by automated responses and limited human review. The lack of clear explanations for either content removal or refusal to act further undermines user trust and discourages reporting.

4.2 Significance and Recommendations for Stakeholders

For online platforms:

- Improve moderation accuracy by combining AI systems with contextual human assessment.
- Increase transparency of moderation decisions through clear and specific explanations to users.
- Collaborate with Roma organizations to develop training materials and context-sensitive content-flagging mechanisms.
- Introduce more detailed reporting categories.
- Build preventive filters for content containing racist symbols, phrases, or visual patterns.

For national institutions:

- Improve coordination between the Commission for Protection against Discrimination, the Ministry of Interior, the Ministry of e-Government, and other competent bodies.
- Establish a national online reporting system that allows real-time case tracking.
- Introduce training programs for public servants on antigypsyism and online hate.

For civil society and educational organizations:

- Develop models for civic digital engagement, including youth monitoring and awareness campaigns.
- Provide emotional support to young people exposed to toxic content.
- Support initiatives that foster dialogue, mutual respect, and joint activities between Roma and non-Roma youth.
- Work toward attitude change through education, media literacy, and exploration of the Stereotype → Prejudice → Discrimination mechanism.

4.3 European Context

Preliminary analyses indicate that online antigypsyism is not confined to a single country but instead follows common European patterns of digital discrimination. The work under TAAO facilitates the development of shared approaches, policies, and tools for response. Final conclusions at the European level will be added after all national reports are completed.

4.4 Limitations

- The sample is non-representative and reflects mainly social media and the networks of the monitors themselves.
- Reporting tools and platform reactions vary, which complicates comparisons.
- Emotional or interpretative biases may occur when decoding ironic or humorous content.

Despite these limitations, the collected data provides a valuable picture of the forms, rhetoric, and consequences of anti-Roma hate online.

4.5 Key Conclusions

The study highlights the need to move decisively from reactive to preventive models for countering hate speech in the digital sphere. While current approaches often rely on post-factum reporting, content removal, or sanctions, these mechanisms address only the visible manifestations of hostility and fail to engage with its deeper social, cognitive, and cultural drivers. Effective prevention requires a holistic framework in which technological tools, platform governance, and public policy are integrated with long-term educational and community-based strategies.

Technical measures—such as improved content moderation systems, algorithmic transparency, and early detection filters—remain important, but they are insufficient on their own. Their impact is limited when users continue to reproduce, share, or passively tolerate harmful narratives. For this reason, preventive efforts must be complemented by comprehensive educational initiatives that strengthen critical thinking, encourage ethical digital participation, and build awareness of how online ecosystems shape perceptions and behaviour.

A central pillar of this preventive approach is the development of digital empathy—the capacity to recognise, understand, and respond to the experiences of others in online environments. Digital empathy counters the depersonalisation and emotional detachment that often enable hate speech to flourish. It supports constructive dialogue, reduces impulsive reactions, and promotes environments where discrimination and dehumanisation are less likely to be normalised.

Equally important is fostering self-awareness regarding the role of personal stereotypes and implicit biases. Many forms of online hostility do not stem from explicit intent but from unexamined assumptions and learned social patterns. Recognising how these biases influence

communication practices is essential for creating sustainable behavioural change. When individuals understand the mechanisms through which stereotypes shape attitudes, they are better equipped to challenge harmful narratives, resist manipulation, and engage responsibly in digital spaces.

Taken together, these elements demonstrate that combating online hate speech is not merely a technical or regulatory challenge but a broader societal process. Prevention must be rooted in education, empathy, and collective responsibility—conditions that empower individuals and communities to build more inclusive, resilient, and safe digital environments.

5. REFERENCES

Alliance against Antigypsyism. (n.d.). *Definition and framework of antigypsyism*. https://www.antigypsyism.eu

Barna, I., & Simonovits, B. (2020). Hate speech and online radicalization: The role of algorithms in amplifying prejudice. *Journal of Digital Society, 5*(2), 21–38.

Bonilla-Silva, E. (2006). *Racism without racists: Color-blind racism and the persistence of racial inequality in America*. Rowman & Littlefield.

Breazu, P., & Machin, D. (2022). *Discourses of racism and nationalism in online media: A multimodal approach*. Routledge.

Council of Europe. (2022). *ECRI General Policy Recommendation No. 15 on combating hate speech*. Council of Europe.

European Commission. (2020). *EU Roma strategic framework for equality, inclusion and participation 2020–2030*. European Commission.

European Commission. (2021). *A more inclusive and protective Europe: Extending the list of EU crimes to hate speech and hate crime* (COM(2021) 777 final).

European Parliament. (2017). *Resolution on fundamental rights aspects in Roma integration in the EU: Fighting anti-Gypsyism (2017/2038(INI))*.

Goffman, E. (1986). *Frame analysis: An essay on the organization of experience*. Harper & Row.

Lentin, A. (2016). Racism in public or public racism: Doing anti-racism in post-racial times. *Ethnic and Racial Studies, 39*(1), 33–48. https://doi.org/10.1080/01419870.2016.1096409

Molnár, E. (2022). Post-positivist grounded theory in qualitative media research. *European Journal of Qualitative Inquiry, 13*(2), 87–102.

PECAO – Peer Education to Counter Antigypsyist Online Hate Speech. (2022). *Final report*. European Roma Grassroots Organisation (ERGO) Network.

Rotaru, I., & Grebeldinger, D. (2025). *Together Against Antigypsyism Online (TAAO): Methodological framework and comparative synthesis report*. Asociatia Nevo Parudimos.

Siapera, E., Moreo, E., & Zhou, J. (2018). *HateTrack: Tracking and monitoring racist hate speech online*. Irish Human Rights and Equality Commission.

Smith, A. (2018). Forms of counter-speech in digital contexts: Ethical considerations and practical approaches. *Journal of Digital Ethics, 7*(3), 45–61.

Smahel, D., Machackova, H., Mascheroni, G., Dedkova, L., Staksrud, E., Ólafsson, K., Livingstone, S., & Hasebrink, U. (2020). *EU Kids Online 2020: Survey results from 19 countries*. EU Kids Online. https://doi.org/10.21953/lse.47fdeqj01ofo

Commission for Protection from Discrimination. (2022). *Annual report on the activities of the Commission for Protection from Discrimination for 2021.* CPD. Retrieved from https://www.kzd-nondiscrimination.com

Ombudsman of the Republic of Bulgaria. (2022). *Annual report on human rights and non-discrimination in Bulgaria – 2021.* Ombudsman. Retrieved from https://www.ombudsman.bg

Center for the Study of Democracy. (2021). *Online propaganda, disinformation and hate speech in Bulgaria.* CSD. Retrieved from https://csd.bg

Bulgarian Helsinki Committee. (2023). *Human rights in Bulgaria in 2022: Annual report.* BHC. Retrieved from https://www.bghelsinki.org

Ministry of Interior. (2023). *Report on hate crimes in Bulgaria for 2022.* Ministry of Interior. Retrieved from https://www.mvr.bg

UNICEF Bulgaria. (2022). *Digital risks, online safety and the experiences of children in Bulgaria: National report 2022.* UNICEF. Retrieved from https://www.unicef.org/bulgaria

Open Society Institute – Sofia. (2021). *Public attitudes toward hate speech and discrimination in Bulgaria.* OSI–Sofia. Retrieved from https://osis.bg

Appendix 1: General monitoring tool

ИНСТРУМЕНТ ЗА МОНИТОРИНГ И АНАЛИЗ НА ПУБЛИКАЦИИ, СЪДЪРЖАЩИ РЕЧ НА ОМРАЗАТА В ИНТЕРНЕТ

I. ОБЩА ИНФОРМАЦИЯ

Социална медия	Тип акаунт				
	Лични/частни акаунти	***Онлайн медийни издания (новини, списания и др.)****	***Институционални публични акаунти*****	***Публични личности*****	***Инфлуенсъри в социалните медии***
Facebook (Meta)					
Twitter (X)					
Instagram					
YouTube					
TikTok					
Друго (моля, уточнете)					

** Официална страница*
*** Правителствена или институционална организация (полиция, Министерство на образованието, държавни агенции и др.), общини, компании, неправителствени организации (включително църкви, училища, културни институции)*
**** Политици, известни професори/учени, артисти, музиканти, журналисти*

Име (на сайта/акаунта/изданието/канала)	
Дата на публикуване:	
Връзка или скрийншот	
Частен или публичен	
Автор (ако има такъв)	

Обхват (брой потребители/достигнати лица):

Харесвания	Нехаресвания (включително „гняв“ или „тъга“)	Споделя ния	Коментари	Преглеждания (Views / Визуализации)

II. ТЕМАТИКА

ОСНОВНА ТЕМА	ОБЯСНЕНИЕ
Престъпления, извършени от роми	
Социални аспекти (жилищни условия, социални помощи, бедност, имиграция и др.)	
Образователни аспекти (отпадане от училище, условия на обучение, стипендии и др.)	
Здравни/санитарни аспекти (пандемия, достъп до болници, аборт и др.)	
Социални движения (протести, граждански права, представителство) и НПО	
Политика (представителство, политически партии, избори)	
Ромски лидери (включително жени)	
Културни събития (музика, филми, театър и др.)	
Спортни събития (състезания, игри)	
Друго (относно роми)	

По тристепенна скала, как оценявате общата рамка на темата:

Позитивно +	Негативно -	Неутрално 0

**Моля, умножете редовете толкова пъти, колкото е необходимо за всяка от идентифицираните теми.*

Тип съдържание	Позитивно +	Негативно -	Неутрално 0
Медийна информация (репортажи, новини и др.)			
Покана (културни/спортни събития, концерти, уебинари и др.)			

Обявления (информация за ежедневни дейности, прессъобщения)			
Гледна точка (редакционен тип, списания, впечатления, коментари, лични мнения)			
Реклама и препоръки (продажби, промоции, работни позиции, пътувания и др.)			
Любопитни факти (специални/необичайни събития, истории)			
Забавление (музика, видеа, филми)			
Други			

Стил на съдържанието	**Позитивно +**	**Негативно -**	**Неутрално 0**
Емоционално			
Формално/Официално			
Призив за действие			
Забавно			
Артистично/фикционално/фантастично			
Научно			
Други			

Съдържа ли визуални елементи?	
Да	**Не**

Моля, посочете използваните нетекстови форми (и техния брой, ако са повече от една)	
Снимки	
Миймове/ GIF-ове	
Карикатури/рисунки	
Мултимедийни материали – reels, story	

Видео	
Анимация	
Други	

По тристепенна скала, как оценявате общата рамка на визуалните елементи:

Позитивно +	Негативно -	Неутрално 0

**Моля, умножете редовете толкова пъти, колкото е необходимо, ако в публикацията има повече от една визуална форма.*

III. ОНЛАЙН ОМРАЗНА РЕЧ – ИНТЕНЗИВНОСТ И НИВА НА ОМРАЗАТА

Форми на явна онлайн омразна реч	**Ниво/степен на омразата**		
	1 приемлива	**2 силна**	**3 екстремно силна**
Груби епитети			
Расистки обиди			
Дехуманизиращи изказвания			
Персонални атаки (ad hominem)			
Наситени емоции (гняв, възмущение, враждебност)			
Подигравки и сарказъм			
Призив към насилие (включително убийство)			
Други			

Форми на закодирана онлайн омразна реч	**Ниво/степен на омраза**		
	1 приемлива	**2 силна**	**3 екстремно силна**
Използване на жаргон (ромски език)			
Ирония			

Псевдонаучни препратки (генетика, фалшиви статистики)			
Расово превъзходство – бял/черен			
Специфична реторика (метонимии, заобиколни изрази, двусмислия)			
Други			

Ориентир:

- ***Използване на жаргон*** *– използване на ромски език/думи за подчертаване на принадлежността към общността.*
- ***Ирония*** *– виж също сарказъм и подигравка („те не могат да отидат в рая, твърде тежки са, за да летят“ ... заради колко бижута са откраднали).*
- ***Псевдонаучни препратки / фалшиви статистики*** *– използване на статистически данни, които не са официални, някои от тях идват от съмнителни научни сайтове или изследвания; например „80% от ромите не искат да работят“.*
- ***Whataboutery*** *(„а какво да кажем за нашите“) – препратка към „нас“ и „тях“, диалектичният общ възглас между „ние“ и „те“.*
- ***Метонимии*** *– изрази като „религия на измамата“, използвани вместо директното назоваване на ромите (без да се споменава думата точно).*
- ***Заобиколни изрази (circumlocutions)*** *– говорене около темата, например: „много мизерия в нашия район“, „толкова тъмно в този блок“.*
- ***Двусмислия (ambiguity)*** *– използване на пунктуация или формулировки, за да се направят реторични забележки, напр.: „трябва ли да бъдат изпратени в концентрационни лагери или не?!“.*

Съдържанието предава ли следните послания? До каква степен?

Статична срещу Динамична онлайн реч на омразата (Letin 2016)	**Степени на токсичност на омразната реч**			
	Пълно несъгласие	**Частично несъгласие**	**Частично съгласие**	**Пълно съгласие**
Затвърждаване на разделението/поляризацията				
Подсилване и затвърждаване на стереотипите				
Разпространяване на митове/заблуди/неистини				
Разпространяване на дезинформация/фалшиви новини				
Оправдаване на изключването,				

стигматизацията и принизяването				
Затвърждаване на ексклузивистки национални идеи				
Други				

Ориентир:

- ***Укрепване на поляризацията*** *= „ние“ срещу „те“*
- ***Затвърждаване на стереотипите*** *= особено негативните (мързеливи, мръсни, неграмотни и др.)*
- ***Разпространяване на митове*** *= свръхестествени сили, магии и др.*
- ***Фалшиви новини*** *= например „те разпространяват болести чрез пътуванията си и начина си на живот“*

IV. ВИДОВЕ КОНТРАИЗКАЗВАНИЯ

1. **Нарушава ли публикацията общоприетите норми на платформата?** Да/Не
2. **Какъв тип действия бихте предприели?**

Начини за противодействие на омразната реч	**Определено да**	**Не съм сигурен**	**Определено не**
Искрено ангажиране			
Игнориране			
Разкриване на тактиката/стратегията на автора			
Оспорване на идеологията, вярванията и ценностите на автора			
Критикуване на аргументите и поведението на автора			
Подиграване или присмиване на съдържанието/на автора			
Отговаряне по същият начин			

Докладване на автора/съдържанието в системата на платформата			
Търсене на правна помощ от страна на гражданското общество			
Докладване на органите/институциите			
Друго			

Ако решите да предприемете контрамерки, моля, предоставете доказателства за вашата дейност:

Вид доказателство	**Дейност - кратко описание**	**Пример – скрийншот или линк**
Текст – коментиране на поста		
Текст – докладване до платформата		
Текст – докладване до институция		
Визуално - скрийншот		

В случай, че има (само) коментари под публикацията/статията/изображението, предоставете някои подробности:

Анализ на коментарите	**Честота**	**Примери (ако има такива)**
Брой коментари		
Брой автори		
Тип на езиково изказване (агресивно vs. мирно) в 5степенна скала	Агресивно = 1 Мирно = 5	
Фалшив профил – скрита самоличност		
Интензитет на типа на коментарите	Липса на омраза = 0 Подигравка = 1 Омраза = 2	

	Подтикване на омразни действия = 3	
Устойчивост: настойчиво коментиране на публикацията с един или повече потребители	Неустойчиво = 0 Слабо = 1 Устойчиво = 2 Интензивно-устойчиво = 3	
Популярност и влияние на негативните коментари в тристепенна скала	Слаба = 1 Силна = 2 Екстремно силна = 3	

V. ДОКЛАДВАНЕ (ЛИЧНИ НАБЛЮДЕНИЯ):
Когато попълвате полето със собствено наблюдение, моля, опитайте се да се позовете на следните подточки (5–6 изречения, включително примери):

- **Негативни констатации за реалния проблем** – обяснение на проблемите/конфликта (обяснени ли са причините; посочено ли е кой е отговорен; има ли тенденция цялата вина да се прехвърля върху ромите);
- **Чии гледни точки са представени** – дали са показани и други гледни точки;
- **Какви стъпки и кога са предприети**;
- **Вашите препоръки** – към кого са насочени (платформата, авторът, институцията);
- **Какъв и кога е получен отговор** (посочете, ако не сте получили отговор);
- **Как е взето решението** – напр. бяха формулирани извинения, публикацията беше изтрита, авторът беше блокиран, авторът беше държан отговорен и др.;
- **Доволни ли сте от отговора/действията** – обяснете защо.

Попълнете полето с личните ви наблюдения за докладвания случай

Основна информация за докладвалия

Страна:	
Име на вашата организация:	
Данните са събрани от:	
Електронна поща:	

Дана на извършване на анализа:	
Дата на подаване на доклада:	

„Заедно срещу антиромските нагласи онлайн“