# Research Report

## WORDS THAT LEAVE SCARS: MAPPING AND UNDERSTANDING ONLINE HATE SPEECH IN NORTH MACEDONIA AND THE WESTERN BALKANS (2022-2025)




Prepared by :

ИНТЕРКУЛТУРА

Co-funded by:

Reporting Diversity Network

Funded by the European Union

March, 2026

# Introduction

Digital communication in the Western Balkans has significantly reshaped the public sphere, with platforms like Facebook, Instagram, TikTok and YouTube becoming central to information sharing, political debate and civic participation. While this has expanded opportunities for expression and inclusion, it has also intensified challenges such as disinformation, polarization and hate speech.

In N. Macedonia, these issues are particularly visible due to the country's multiethnic and socially complex context. Online spaces often reflect and amplify existing social tensions, with hate speech targeting vulnerable groups such as Roma communities, women, persons with disabilities and LGBTQI+ individuals. This form of communication is frequently normalized through humor, irony or claims of "free speech", allowing it to spread widely.

The visibility of hate speech has increased, especially during politically or socially sensitive periods. It contributes not only to online hostility but also to real-world discrimination, mistrust and social division. As a result, online hate speech represents a broader human rights and social cohesion issue.

This report combines Meta's transparency data (2023-2025) with independent monitoring by Interkultura Skopje to provide a comprehensive analysis of hate speech trends in N.Macedonia. By integrating quantitative data with qualitative insights, the research aims to better understand how hate speech operates, who it affects and how it can be addressed.

Ultimately, the project "Words that leave scars" seeks not only to document the problem but to support more inclusive, responsible and empathetic digital communication in N. Macedonia and the Western Balkan

# Research Objectives




The central objective of this research is to identify and analyze patterns of online hate speech in N.Macedonia, situating them within the broader regional and global context of digital communication. Hate speech, as a linguistic and social phenomenon, does not exist in isolation, it reflects deeper power structures, social hierarchies and political dynamics. Understanding how hate manifests, circulates and is perceived online requires an integrated approach that combines data-driven analysis with cultural and ethical interpretation.

This study was therefore conceived with the intention of bridging quantitative evidence and qualitative insight merging systematic monitoring of digital platforms with the lived experiences of individuals and communities who are most affected by online hostility. The research seeks to clarify the different kinds of hate speech that exist. It also explores why these forms of speech persist, how they change over time and what effect they have on social cohesion and public discussion in N.Macedonia.




The research aims to outline the extent, types and targets of online hate speech from 2022 to 2025. This involves examining the prevalence of hate-based communication across various digital platforms including Facebook, Instagram, TikTok, YouTube and online news portals and classifying it according to target groups such as Roma communities, women, persons with disabilities and LGBTQI+ individuals. By tracking the frequency and intensity of these narratives, the study provides an empirical overview of the dominant patterns of exclusion and hostility in the country's online ecosystem.

Another important part of the research is combining independent monitoring data with insights from the platform. This includes the Meta transparency dataset for 2023-2025. This collaboration provides two viewpoints on hate speech: one from civil society observation and another from platform-level data on content moderation. By comparing the data collected by Interkultura and RDN with Meta's enforcement statistics, the research looks at the link between public behavior, platform algorithms, and regulatory responses. It evaluates how well automated systems and user reporting tools find and reduce hate speech. It also points out the shortcomings of algorithmic moderation in culturally specific areas like the Western Balkans.

# Research Objectives

A third layer of the research focuses on the social and psychological impact of online hate speech. Beyond its measurable digital footprint, hate speech carries profound emotional and symbolic consequences. It shapes how individuals perceive themselves and others, how communities are represented in media discourse and how belonging is negotiated in the digital public sphere. The study thus seeks to understand how online narratives influence self-perception and representation among marginalized groups particularly youth, who are both the most exposed to and most capable of challenging these harmful discourses.

This research aims to create useful knowledge that can guide future policy, education and media practices. The main goal is not just to describe or measure hate speech. It seeks to contribute to solutions by identifying key areas for action, recommending evidence-based policy measures and supporting initiatives that promote inclusive communication. The findings are meant to assist journalists, educators, civil society organizations and decision-makers in developing lasting strategies to prevent, counter and turn hate speech into dialogue and understanding.

In this way, "Words that leave scars" positions research as both a diagnostic and transformative tool a means to document the realities of digital hostility while laying the intellectual foundation for a more empathetic and equitable media culture in N.Macedonia and the wider Western Balkan region.



# Methodology

## Research data

This research applies a mixed-method research design combining quantitative monitoring of digital content with qualitative data gathered through participatory workshops and focus group discussions. The methodology was designed to provide both statistical insight into the scale of online hate speech and deeper understanding of how such narratives are perceived and experienced by young people and marginalized communities in N.Macedonia. By integrating platform-level transparency data, civil society monitoring and participatory qualitative research, the study ensures both analytical depth and contextual relevance.

The quantitative dimension focused on systematic content analysis of social media posts, comments and media articles published between 2022 and 2025. This provided a measurable picture of how hate speech manifests across digital platforms, what groups are most frequently targeted and how such content circulates online.



# Research data

The qualitative dimension complemented this by engaging directly with the human experiences behind the data. Through semi-structured interviews and focus group discussions, the research explored how online hate is perceived, internalized and resisted by individuals and communities most affected by it. This allowed the study to move beyond numbers and categories, highlighting the emotional and social consequences of hate speech, as well as the strategies people use to navigate digital hostility.

Finally, the research included a platform-level analytical layer by integrating Meta's ransparency and DSA compliance Reports (Meta dataset). These official datasets, released under the EU's Digital services act framework, provide detailed country-level statistics on content moderation. This includes removals, user reports and appeals related to hate speech on Facebook and Instagram. Their inclusion allows for a triangulated analysis. It links observations from the ground level with enforcement methods on the platform. It also shows possible gaps between what policies promise and what really happens.

These three components quantitative, qualitative and platform-based form a holistic methodological framework that enables both critical and evidence-driven analysis of online hate speech trends in N.Macedonia.



# Data sources

To ensure validity and comparability, the study relied on three primary categories of data:

1. Primary dataset: The core dataset was produced through systematic monitoring conducted by Interkultura's research team as part of the. It consists of more than 3.000 public social media posts, comments and online media articles that were manually coded according to standardized hate-speech indicators. The sample spans multiple digital environments from Facebook and Instagram to TikTok, YouTube and online news portals covering both individual user-generated content and institutional communication by public figures or media outlets.
2. Meta transparency data: This dataset was obtained from Meta's official Digital services act (DSA) transparency reports, released for the European Union and associated markets, including the Western Balkans. It contains aggregated data on hate-speech removals, user reports, appeals and moderation coverage for N.Macedonia and neighboring countries. The Meta data enable comparative insight into how major platforms detect and act upon hate-speech violations, offering a rare quantitative view into content moderation practices.
3. Secondary data: To complement primary data, the research incorporated a wide range of secondary sources, including reports and studies by ILGA-Europe, BIRN, OSCE, the U.S. State Department and academic research on digital hate speech in Southeastern Europe. These sources were crucial for contextualizing findings within regional and international frameworks and for identifying recurring patterns across time and geography.

# Data collection and sampling

Data collection followed a structured and replicable process designed to maximize representativeness and minimize bias. The monitoring team focused on publicly accessible digital spaces, prioritizing content with significant visibility and engagement potential.

The research team systematically sampled content from:

- Facebook pages and groups associated with media outlets, influencers, public officials and community networks
- Instagram and TikTok accounts that have a large youth following or are recognized for influencing public discussions
- YouTube channels and podcast comment sections, where longer-form discussions often produce hate-laden narratives
- Major online news portals, especially those with open comment sections or partisan editorial framing.

Each piece of content was coded by three trained researchers according to the following variables:

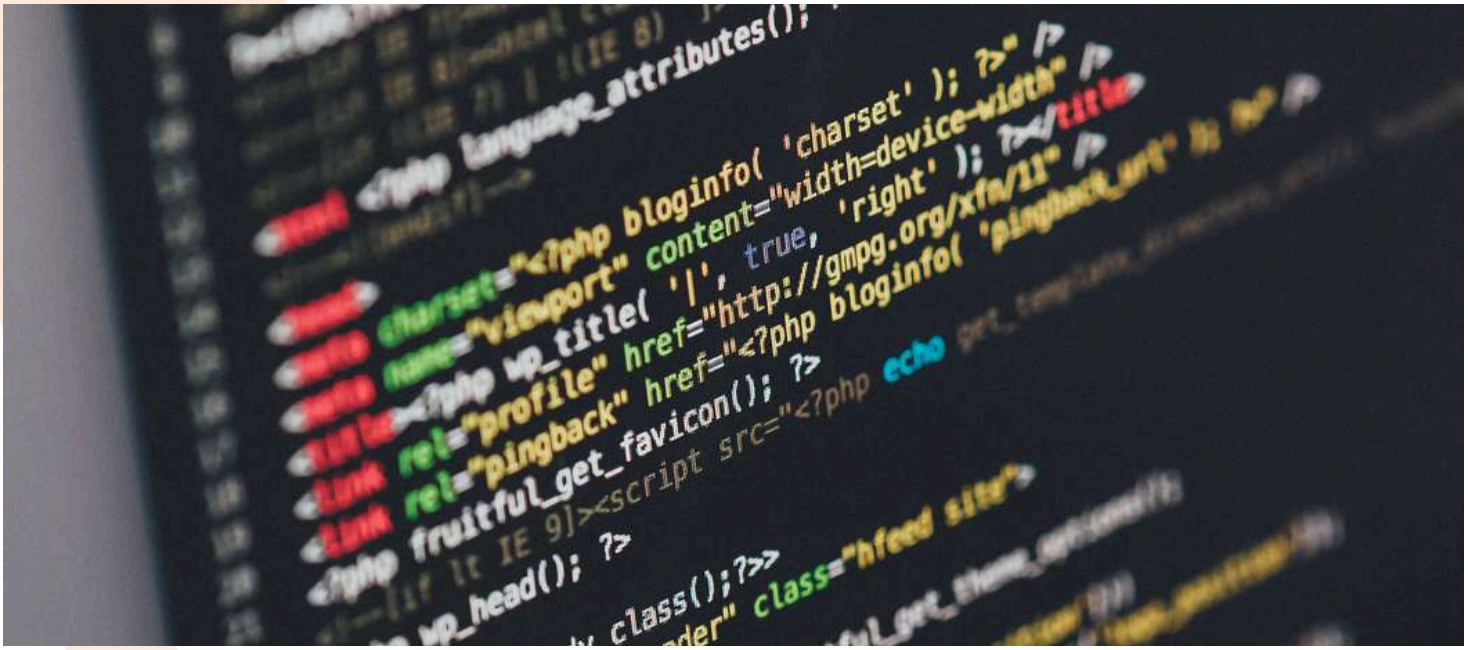
- Platform and source type (social media / news portal)
- Author identity (private individual, influencer, journalist, politician, or organization)
- Target group (ethnicity, gender, disability, political ideology, sexual orientation, etc.)
- Type of hate speech (insult, threat, dehumanization, mockery, stereotyping, etc.)
- Engagement level (likes, shares, comments)
- Tone and format (text, meme, video, mixed content)

The research team worked independently but followed a shared coding manual to ensure consistency.

This process created a strong quantitative dataset that can identify both the amount and nature of online hate speech. It also allows for cross-tabulation by platform, target group and narrative type.

# Meta data integration

The integration of Meta data represented a key innovation in the research methodology, enabling direct comparison between civil society monitoring and platform-level enforcement.



The dataset, obtained through Meta's publicly accessible transparency reports and specialized DSA repositories, provided national-level insights into content moderation across Facebook and Instagram.

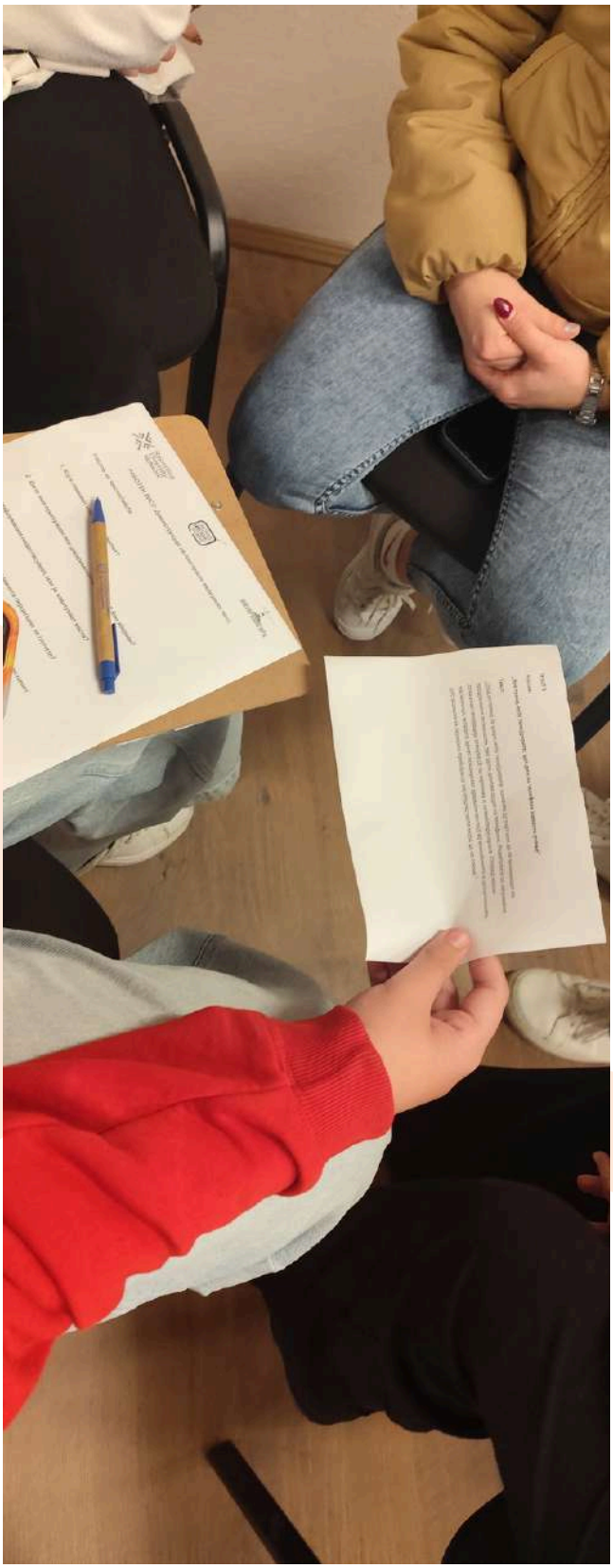
The following key indicators were analyzed:

- Total hate-speech removals (Q1 2023 - Q2 2025): approximately 38.000 items in N.Macedonia, of which 94% were detected automatically through machine-learning systems.
- User reports: around 7.500 reports submitted by users in the same period, with 69% resulting in enforcement action (content removal or warning).
- Appeals: approximately 480 appeals were lodged by users whose content was removed 23% of these were reinstated after review, indicating potential overreach in moderation.
- Regional moderation coverage: 88% of hate-speech detection across Macedonian, Albanian and Serbian languages combined, with remaining gaps primarily in dialectal or coded language variations.

These figures were checked against the Interkultura dataset to find any discrepancies and confirm trends. The analysis showed that increases in Meta's removal statistics matched closely with politically sensitive times, like the 2024 national elections and discussions on gender and minority rights. This connection boosts confidence in the research's validity and in the importance of platform transparency data as a useful analytical tool.

# Qualitative component

While quantitative monitoring provides insight into the scale and distribution of hate speech online, qualitative research is essential for understanding the meanings, perceptions and emotional consequences behind these patterns



For this reason, the research incorporated a qualitative component based on participatory workshops and focus group discussions with young people.

Between October 2025 and March 2026, a total of five in-person focus group sessions were organized in five cities across North Macedonia: Bitola, Kavadarci, Kumanovo, Veles and Skopje. In addition, one online discussion session was organized in order to include participants who were unable to attend the in-person meetings and to expand the geographical coverage of the study.

These sessions were implemented within the framework of the project “Words that leave scars”, which combines research with youth education and public awareness activities aimed at addressing hate speech in digital media environments. The qualitative sessions were integrated into a series of educational workshops focused on media literacy, human rights and critical engagement with online narratives.

In total, approximately 80 youth participants were involved in these activities, including around 60 participants in the in-person workshops and focus groups and 20 participants in the online discussion group. Participants included high school and university students, youth activists and young media consumers from diverse social backgrounds.








Each focus group followed a semi-structured discussion format facilitated by trained moderators from the Interkultura project team. The sessions lasted between 90 and 120 minutes and explored several thematic areas, including:

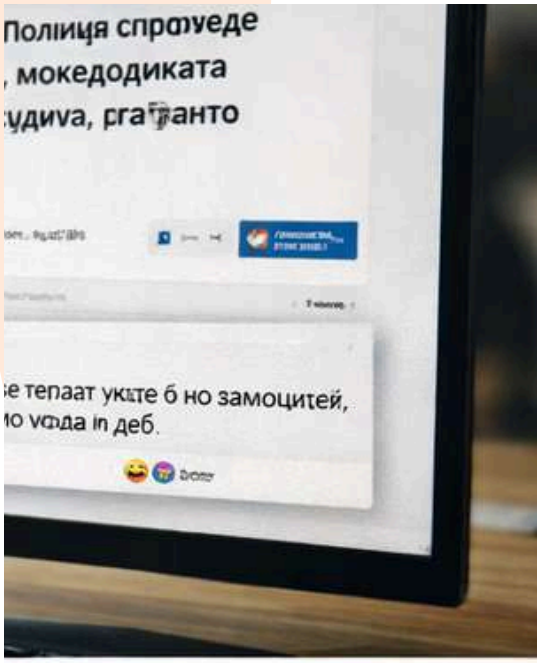
- participants' everyday exposure to hate speech on social media platforms
- perceptions of how marginalized communities are represented in digital media
- emotional and psychological responses to discriminatory language online
- participants' willingness and ability to report or challenge hate speech
- ideas for creating positive counter-narratives and promoting inclusive communication.

Participants were encouraged to share examples from their own online experiences, particularly from widely used platforms such as Facebook, Instagram, TikTok and YouTube. These discussions allowed the research team to identify recurring narratives, social attitudes and behavioral patterns related to hate speech among young digital media users.

All sessions were conducted in accordance with ethical research standards. Participants were informed about the purpose of the study, and discussions were documented through facilitator notes and anonymized summaries to ensure confidentiality.

The qualitative data collected through these focus groups complement the quantitative monitoring results by providing deeper insight into how hate speech is interpreted, normalized and challenged within everyday online communication. They also highlight the potential role of youth as agents of change in promoting more responsible and inclusive digital discourse.

# Quantitative findings






## *Overview: North Macedonia (2022-2025)*

The quantitative analysis conducted as part of the "Words that leave scars" research provides a detailed statistical overview of online hate speech dynamics in N.Macedonia over a three-year period, from 2022 to 2025. The monitoring process identified 34 high-impact incidents of hate speech that gained notable visibility across social and digital media platforms. These cases represent not isolated acts of hostility, but recurring communication patterns that reflect broader social anxieties, stereotypes and prejudices embedded within the country's online discourse.

Although online hate takes many forms, such as direct verbal attacks, mocking memes, coded language and sarcastic "humor," the data show consistent patterns in the targeted groups and the platforms where this content thrives. The distribution of targets closely matches regional trends noted by the Reporting Diversity Network (RDN) and platform-level data from the Meta transparency dataset. This highlights that the issue of hate speech in N. Macedonia is part of a broader systemic pattern across the Western Balkans.

# Cross-validation with Meta data




The integration of Meta transparency data (2023-2025) provides an important comparative layer that strengthens and contextualizes these findings. According to Meta's transparency metrics, the volume of hate-speech removals peaked in mid-2024, corresponding with the pre-election period in N.Macedonia. This surge reflects the politicization of public discourse, where ethnic, gendered and ideological narratives were frequently weaponized by online actors to mobilize supporters or discredit opponents.

Meta's data show that approximately 38.000 pieces of content were removed from Facebook and Instagram for hate-speech violations during the analyzed period, with 94% of these removals resulting from automated detection systems rather than user reports. Only 7.500 items (19%) were flagged directly by users, of which 69% resulted in action. This disparity reveals a telling insight: citizens are not actively reporting hate speech, suggesting that hostility and discrimination have been partially normalized in digital communication.

Furthermore, 480 appeals were filed by users whose content had been removed and in 23% of cases, Meta reinstated the content indicating that automated moderation occasionally misclassifies context-dependent or sarcastic speech as hate. While automation ensures efficiency, these figures confirm that algorithmic moderation struggles with cultural nuance and linguistic specificity, particularly in smaller language environments such as Macedonian and Albanian.

These insights reinforce the argument that platform data alone cannot capture the full reality of online hate. Civil society monitoring remains essential for identifying local idioms, coded language and implicit narratives that evade detection by machine-learning filters. The combination of Meta transparency data and Interkultura's independent dataset therefore creates a more balanced picture one that acknowledges both the scale of the problem and the limits of current technological solutions.

# Regional perspective




**When placed within the broader Western Balkans context, the findings from N.Macedonia fit into a regional pattern of rising polarization and online hostility.**

According to the Regional Monitoring Report (May 2022-December 2023), a total of 523 cases of hate speech were recorded across Albania, Bosnia and Herzegovina, Kosovo, Montenegro, N.Macedonia and Serbia. The regional distribution highlights similar trends in target groups:

- Ethnicity: 28,1% of all recorded cases
- Gender: 24,4%
- Political / ideological: 14,8%
- Sexual minorities: 13,6%
- Other vulnerable groups: 19.1% (including migrants, refugees and people with disabilities)

During this same period, Meta reported approximately 310.000 hate-speech removals across the Western Balkans region, with 91% of all takedowns performed automatically and only 2% of user appeals upheld. This confirms that large-scale moderation systems are in operation, but that their effectiveness remains limited by context and precision.

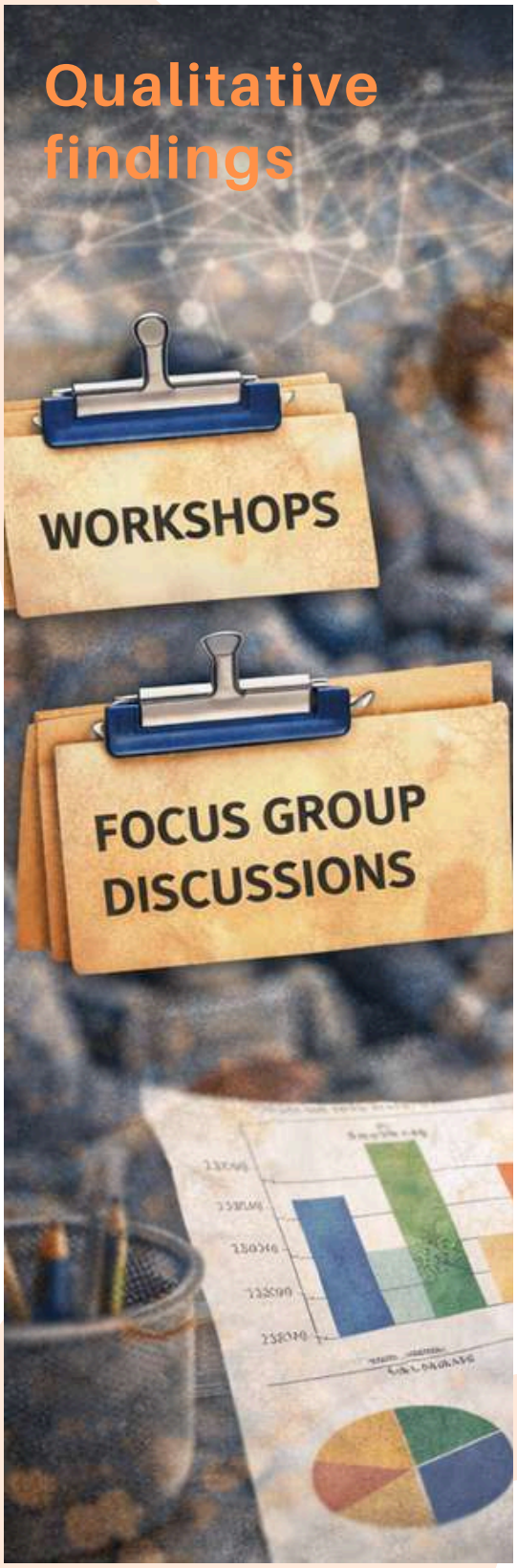
The consistency between regional monitoring and Meta's aggregated enforcement data illustrates a shared challenge across the Western Balkans: while hate speech is widespread and regularly identified by algorithms, its underlying social drivers such as inequality, nationalism, and patriarchy, remain unaddressed. As a result, technological moderation, while necessary, treats only the surface symptoms of a much deeper societal issue.

The quantitative findings presented here reveal an alarming yet clear picture: hate speech in N.Macedonia is not an isolated digital phenomenon, but part of a broader regional and cultural structure of hostility and exclusion. The data, when viewed together with platform-level information, show that hate thrives where public accountability is weak and digital empathy is scarce. These insights set the stage for the next chapter of this research a qualitative exploration of the narratives, emotions and lived experiences that give shape to this ongoing digital crisis.

| Target group | Share (%) | Key platforms | Common narrative |
|---|---|---|---|
| Women/Gender | 32% | YouTube, Facebook | Misogyny, trivialization of violence, sexualized insults toward women in media and politics |
| Ethnic groups (Roma, Albanians,Macedonians ) | 26,5% | Facebook, X (Twitter) | Racial slurs, exclusion rhetoric, framing minorities as "outsiders" |
| Political / Ideological | 14,5% | TikTok, Facebook | |
| LGBTQI+ Individuals | 12% | Instagram, Facebook | "Traditional values" vs. equality narratives, dehumanizing humor around Pride events |
| Persons with disabilities | 8% | Facebook | Mockery, infantilization, use of disability as metaphor for incompetence |
| Other minorities (Refugees, Migrants) | 7% | Mixed platforms | Fear-based and nationalist narratives portraying refugees as a threat |

The data indicate that gender-based and ethnic hate speech remain the dominant categories, together accounting for nearly 60% of all identified cases. Women, particularly those in public or professional roles, were subject to targeted misogyny often expressed through demeaning comments, sexualized slurs, or ridicule aimed at undermining their credibility. Meanwhile, Roma and Albanian communities faced continuous ethnic stereotyping, typically framed through language that portrayed them as either "lazy," "criminal," or "unfit to belong."

Facebook and YouTube emerged as the primary platforms for dissemination, largely because of their wide reach and permissive comment structures, which enable rapid viral spread of hateful or sensational content. TikTok, while newer, is quickly becoming an influential arena for political and ideological hostility, especially among youth audiences. The rise of short-form video content as a carrier of hate narratives marks a significant shift from text-based insults to visualized, performative forms of hate that are harder to moderate.




## *Dominant narratives*

The qualitative phase of the "Words that leave scars" research offered a crucial human dimension to the quantitative data, providing insight into the ways hate speech is experienced, understood and resisted by those most affected by it. Through a series of semi-structured interviews with journalists, youth activists and educators and focus group discussions with members of Roma, women's and disability-rights communities, the research captured voices that are too often excluded from public conversations about media and social cohesion.

What emerged from these discussions is a layered picture of digital life in N.Macedonia one in which the internet functions simultaneously as a tool of empowerment and a space of constant exposure to hostility. Participants described an online environment saturated with negativity, where humor and irony often serve as camouflage for prejudice and where the absence of accountability allows discriminatory narratives to spread unchecked. The following subsections unpack the most prominent themes that surfaced from these conversations.

# Normalization of Hate through Humor




Across all focus groups, participants noted that hate speech is frequently disguised as "just a joke." Memes, GIFs and short TikTok videos use humor to trivialize discrimination, particularly against women and Roma communities. Participants emphasized that such "humor" often carries the same emotional weight as direct insult, but is harder to challenge publicly because it hides behind irony.

A young woman from the Skopje focus group described it poignantly:

"When someone says something sexist or racist and then adds "Relax, it's a joke", it becomes impossible to argue. If you react, you're called oversensitive, if you stay silent, it keeps spreading."

This normalization of hate through comedic framing reflects a deeper social mechanism a way of maintaining prejudice while disavowing responsibility for it. The digital environment, with its rapid-fire content and emphasis on engagement, rewards provocation and sarcasm, creating what several participants called "an economy of outrage."



# Moral Panic and the Framing of "Tradition"

Another recurring narrative centered on what interviewees called "moral panic." Online debates often frame marginalized groups as threats to national identity, morality, or "traditional values." This rhetoric is especially visible during public discussions of gender equality or LGBTQI+ rights. Posts and comments frequently portray progressive movements as "foreign imports" or "Western agendas," framing equality and diversity as cultural invasions rather than human rights issues.

A Roma youth activist reflected:

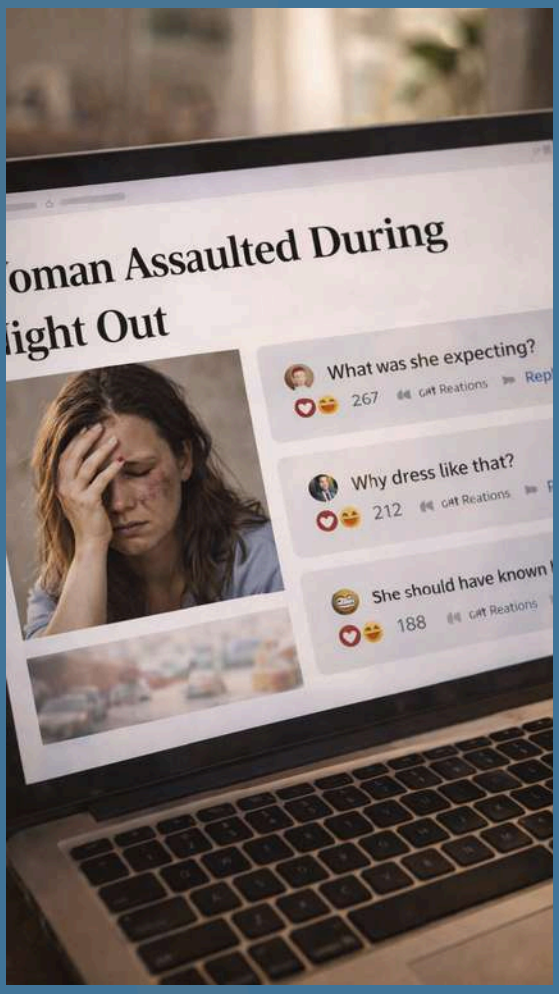
"Every time there's talk about diversity or inclusion, someone online says we're destroying the culture, that this is not who we are. They make it seem like equality is an attack on the nation."

This moral-panic framing creates a symbolic battle between "us" (the defenders of tradition) and "them" (those demanding change). It reinforces an exclusionary nationalism that conflates social conservatism with patriotism, legitimizing hate as a form of cultural protection.






# Victim Blaming and the Reversal of Responsibility




A particularly harmful narrative identified in both interviews and focus groups involves victim blaming the idea that marginalized individuals somehow provoke the hate they receive. Women who speak publicly or express feminist views online reported being told that their visibility "invites" harassment. Similarly, Roma participants noted that when they respond to hate speech, they are often accused of "playing the victim."

One journalist interviewed summarized this dynamic: "The logic is always twisted. When women get harassed, people ask why they posted that photo. When Roma people respond to racism, they're told they're too sensitive. It's as if the problem is not the hate itself, but our reaction to it."

This reversal of responsibility not only silences victims but also reinforces the perception that hate speech is an inevitable part of online life, something to endure rather than challenge.



# Political Opportunism and the Instrumentalization of hate

Several participants, including journalists and NGO representatives, spoke about the political dimension of online hate. Election periods and politically charged moments often see spikes in divisive narratives, with politicians and party-affiliated media using hate speech directly or indirectly to energize supporters and discredit opponents.

One participant described Facebook comment sections during election season as "digital battlefields," where insults and ethnic stereotypes are strategically circulated to mobilize voter emotion. In this sense, hate becomes not just a byproduct of public discourse but a deliberate political tool.

This instrumentalization of hate reveals a troubling continuity between the online and offline spheres: the rhetoric of division that dominates politics is mirrored, amplified and distorted in digital spaces, shaping a social climate of constant antagonism.






# The Impact on Targeted Groups



**The emotional and psychological impact of online hate was a central focus of the qualitative inquiry.**

Participants across all groups described feelings of anxiety, exhaustion and self-censorship, resulting from continuous exposure to hostile environments. Many respondents stated that they had reduced their online activity or avoided posting personal opinions out of fear of ridicule or harassment.

A young woman participant explained:
"I used to comment on social issues, but after being attacked a few times, I stopped. It's not worth the stress. You start questioning yourself am I the problem for speaking up?"
This sentiment of withdrawal was echoed by Roma and disability-rights activists, who emphasized that hate speech not only discourages individual expression but also erases entire communities from public visibility. As one Roma participant said,
"The media only talk about us when there's hate never when we do something good."

Participants also discussed the psychological toll of constant exposure to degrading language. Women and Roma participants described a sense of double vulnerability being targeted both for gender and ethnicity while persons with disabilities spoke of the pain of infantilization, where online users dismiss their autonomy or mock their conditions for entertainment.

These experiences collectively illustrate how hate speech extends beyond verbal aggression. It produces structural silence discouraging participation, limiting representation, and reinforcing social hierarchies. The result is a cycle of exclusion: those most affected by hate are driven out of the digital spaces where public opinion is formed, leaving those spaces dominated by the loudest and most aggressive voices.

# The Paradox of visibility



**An emergent theme across discussions was what participants described as a "visibility paradox."**

For marginalized individuals, being visible online can be both empowering and dangerous. Visibility enables advocacy, storytelling and community building, but it also attracts disproportionate levels of scrutiny and hate.

As one disability-rights advocate put it:
"The moment you speak about inclusion or rights, you become a target. But if you stay silent, people say we don't exist."

This paradox illustrates a key dilemma of the digital age: the struggle to be seen and heard without becoming exposed to harm. Participants expressed a collective desire for safer online spaces where engagement would not come at the cost of dignity or emotional well-being.

In summary, the qualitative findings reveal that hate speech in North Macedonia operates not merely as offensive language, but as a cultural and psychological system, one that normalizes exclusion, legitimizes inequality and silences those who most need to be heard. Through humor, moral panic, victim blaming and political manipulation, hate becomes woven into everyday communication, shaping both self-perception and public discourse.

These narratives and their emotional consequences highlight the urgent need for education, empathy-building and responsible media practices. The stories shared in this research remind us that the scars left by words are real and that understanding their impact is the first step toward healing them.

# Discussion

The integration of Meta's enforcement data with independent civil-society monitoring reveals a complex and often contradictory reality a dual landscape in which progress and neglect coexist. On one side, global platforms like Meta have made measurable strides in detecting and removing hateful content through automated systems and transparency reporting. On the other, local experiences show that these same systems often fail to grasp the nuances of language, culture, and intent that define hate speech in smaller linguistic and socio-political contexts such as N.Macedonia.

Meta's transparency reports provide an encouraging image of action: tens of thousands of hate-speech removals, consistent algorithmic surveillance and compliance with European regulatory frameworks under the Digital services act (DSA). However, when examined alongside Interkultura's independent monitoring and the testimonies gathered from affected individuals, a more nuanced and troubling picture emerges. Automated moderation captures volume, not meaning. It identifies hate as a statistical pattern a word, a phrase, a trigger without necessarily understanding its social and emotional significance.

The consequence of this gap is that many forms of hate speech remain invisible to the algorithms. Sarcasm, coded language and context-specific insults, often embedded in local dialects or cultural references pass undetected, allowing discrimination to circulate freely under the guise of humor or "freedom of speech." Participants in the focus groups consistently reported encountering hate that did not meet the "threshold" of removability: it was not explicit enough to be flagged by platforms but was clearly harmful in its tone and subtext. This creates a false sense of safety within digital spaces, one where platforms can claim success in enforcement while communities continue to experience hostility on a daily basis.

# Discussion

Equally concerning is the low rate of user reporting revealed by Meta's data. Despite the prevalence of hate content, only a small fraction of users actively report it. The reasons are multifaceted: fear of online retaliation, lack of trust that reporting mechanisms will result in meaningful action and perhaps most disturbingly the normalization of digital aggression. Many citizens, especially younger users, perceive hate speech as an inevitable part of online interaction, something to be tolerated rather than challenged. This normalization erodes empathy and reshapes digital citizenship into passive consumption rather than active responsibility.

This situation points to a deeper systemic issue: the absence of shared ownership in creating safe online spaces. Moderation, when left solely to corporate algorithms, becomes detached from local realities. Conversely, when civil society efforts remain isolated, they lack the scale and technical capacity to meaningfully alter the digital environment. Bridging this divide requires collaborative moderation models, where technology companies, researchers and community organizations work in tandem. Civil society actors possess the contextual and linguistic knowledge that global platforms often lack, while platforms hold the technical tools and reach needed to enforce standards at scale.

Such collaboration could take multiple forms, from joint training programs for content reviewers in local languages, to shared early-warning systems where CSOs flag harmful trends before they escalate, to partnerships that feed anonymized datasets into machine-learning systems to improve cultural sensitivity. These community-based moderation frameworks would allow detection algorithms to be informed by human understanding by the subtleties of local culture, idiom and emotion that machines alone cannot interpret.

# Discussion

Furthermore, the findings suggest that education and digital literacy must become central pillars of any sustainable solution. Encouraging users to recognize hate speech, understand its consequence and use reporting tools effectively would not only increase the accuracy of moderation but also rebuild public confidence in online participation. When citizens perceive that their actions can make a difference, they are more likely to engage responsibly and challenge hate rather than silently endure it.

In essence, this discussion reveals that combating hate speech is not merely a technological or regulatory task it is a profoundly social and ethical project. Algorithms may remove words, but they cannot heal wounds. True progress will come only when platforms, policymakers and communities act together to transform digital culture from one of hostility and exclusion into one of empathy and accountability.

The dual reality uncovered by this research where automated systems operate efficiently yet meaningfully fail underscores the urgency of building a shared ecosystem of digital responsibility. It is within this collaborative framework that N.Macedonia and the wider Western Balkans can begin to reclaim online spaces as arenas of dialogue, inclusion and dignity rather than division and fear.

# Key findings

The findings of this research reveal that online hate speech in N.Macedonia is both pervasive and deeply normalized within the country's digital culture. Far from being an isolated or spontaneous phenomenon, it has become a structural feature of online communication, shaped by long-standing social hierarchies, political polarization and cultural prejudices. The data collected through both independent monitoring and Meta's transparency reports point to the same conclusion: the problem is widespread, multifaceted and sustained by a system that privileges engagement over empathy. Insights from the youth focus groups conducted across five cities in North Macedonia further confirmed that exposure to online hate speech is widespread among young social media users, many of whom perceive discriminatory language as a normalized part of online interaction

Gender-based and ethnic hate remain the most visible and persistent forms of online hostility. Women, particularly those who participate in public discourse or hold professional roles in media, politics, or activism, continue to be targeted through sexualized insults, objectification and moral judgment. Ethnic minorities, especially Roma and Albanian communities, are framed through a discourse of exclusion and inferiority that portrays them as burdens or outsiders. These narratives are not confined to fringe groups but often appear in mainstream comment sections, meme pages and influencer content. The persistence of such language suggests that online hate reflects, rather than creates, social inequality. The digital sphere functions as an amplifier of biases that already exist in society.

The research also shows that social media platforms, particularly Facebook and TikTok, play a central role in the dissemination of hate speech. Their algorithms, designed to maximize visibility and interaction, tend to reward emotionally charged and divisive content. As a result, hateful messages often spread faster and wider than positive or neutral information. This dynamic, combined with weak moderation and limited public awareness, creates an environment where hate speech not only survives but thrives. The rise of short-form video content on TikTok and Instagram reels introduces a new layer of complexity, as discriminatory messages are now encoded through humor, performance and visual cues rather than explicit verbal statements. Such forms of communication are more difficult to detect through traditional moderation systems and often remain publicly accessible for long periods.

# Key findings

Another key finding concerns the discrepancy between platform-level data and lived experiences. While Meta's reports indicate significant progress in automated detection and content removal, the testimonies collected from affected individuals tell a different story. Participants continue to encounter hate on a daily basis, suggesting that the metrics of success used by large platforms do not adequately reflect the quality of online experience for marginalized users. Automated systems excel at identifying explicit slurs or repeated keywords but often fail to recognize coded language, irony, or cultural references that carry the same discriminatory intent. The research therefore highlights a gap between quantitative enforcement and qualitative reality.

The low rate of user reporting further reinforces this gap. Despite the scale of online hostility, only a small percentage of users actively report hate speech when they see it. Interviews and focus groups revealed several reasons for this reluctance: fear of online backlash, lack of confidence in the effectiveness of reporting mechanisms, and a general sense of resignation that digital aggression is simply part of the online environment. This normalization of hate undermines civic participation, as many individuals, especially those from marginalized communities choose withdrawal and silence over confrontation. In doing so, the public sphere becomes dominated by voices that are loudest rather than most representative.

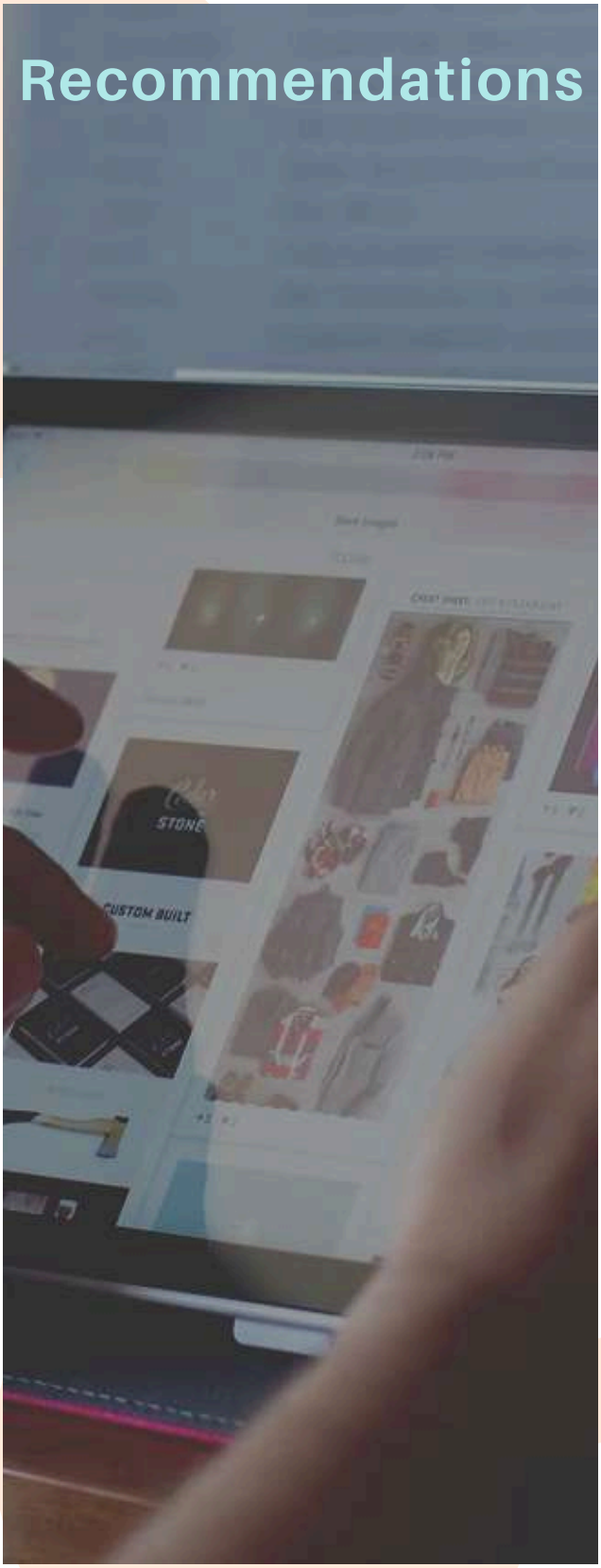
A particularly striking insight emerging from the research is the connection between political cycles and spikes in hate speech. Periods surrounding elections or public debates on gender and minority rights often coincide with noticeable increases in online hostility. Politicians and partisan media outlets, intentionally or not, contribute to these dynamics by using inflammatory rhetoric that filters down into social networks. In such moments, digital spaces become arenas of emotional mobilization rather than informed debate, where misinformation and hate thrive on political polarization. This cyclical pattern underscores that online hate speech is not merely a matter of personal behavior but a reflection of institutional and political conditions.

# Key findings

The human consequences of these digital dynamics are profound. Participants described feelings of fear, exhaustion and alienation. Women spoke about withdrawing from online platforms after repeated harassment. Members of the Roma community emphasized how constant exposure to racist language reinforces their social invisibility. Persons with disabilities described being treated online not as equals but as symbols of pity or ridicule. These experiences reveal that the harm caused by hate speech extends far beyond the screen. It affects mental health, shapes identity, and determines who feels entitled to speak and who feels forced into silence.

In essence, the study demonstrates that the online environment in N.Macedonia mirrors the inequalities of the offline world. Hate speech operates as both a symptom and a driver of social exclusion. While digital technologies have created new opportunities for connection, they have also given old prejudices new platforms and greater reach. The data and testimonies collected through “Words that leave scars” confirm that combating hate speech requires more than technical moderation. It demands cultural change, education and the rehumanization of communication itself. Only when technology, policy, and community engagement converge can digital spaces begin to reflect the values of respect, empathy and equality that a democratic society aspires to uphold.

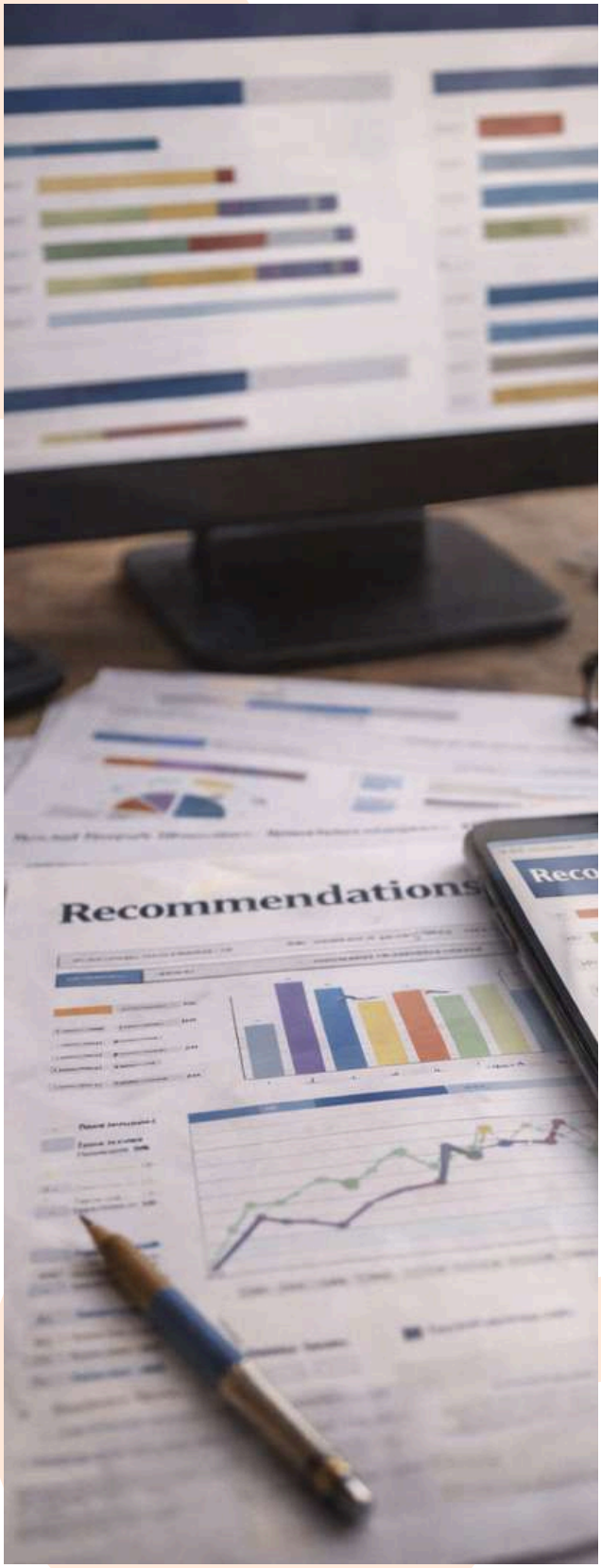
# Recommendations



The findings of this research clearly indicate that online hate speech in N.Macedonia is a systemic issue that requires a coordinated and multi-level response. It cannot be addressed solely through stricter moderation or legal frameworks. Rather, it demands a holistic approach that combines technological accountability, civic education, media reform and long-term cultural change. The following recommendations are therefore presented as interconnected pathways rather than isolated solutions. They are meant to be understood as part of a broader ecosystem of responsibility shared by platforms, policymakers, journalists, educators and civil society organizations.

For social media platforms such as Meta, TikTok and YouTube, the most urgent task is to strengthen their understanding of local contexts and languages. Automated systems can identify keywords and patterns, but they fail to grasp the nuance of regional dialects, irony, or coded hate expressions that often carry discriminatory meaning in subtle ways. To overcome this limitation, platforms should collaborate directly with local civil society organizations and researchers who possess cultural and linguistic expertise. Such cooperation could take the form of advisory partnerships, shared datasets and joint monitoring initiatives that feed context-sensitive examples into machine-learning algorithms. Platforms should also improve transparency by providing disaggregated, country-level data on content moderation outcomes, including how many appeals are upheld, how long harmful content remains visible and what languages are covered by automated systems. These steps would not only enhance detection accuracy but also rebuild public trust in the fairness and accountability of moderation processes.

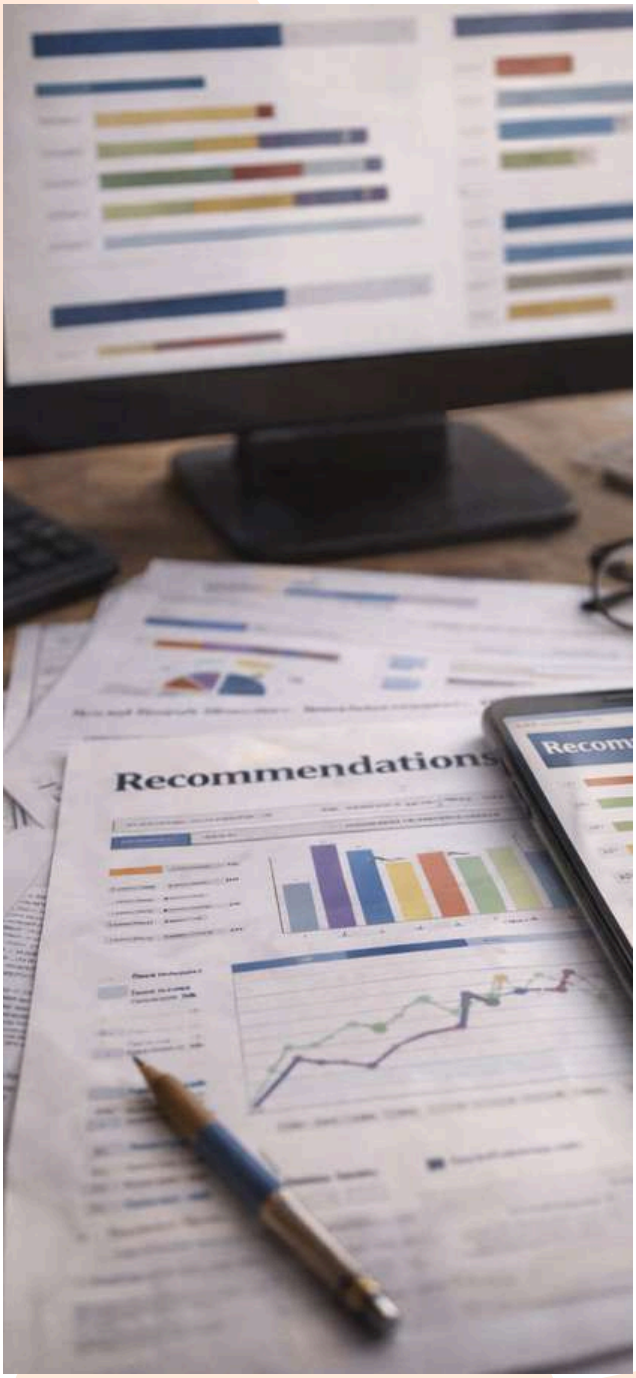
# Recommendations




For policymakers and regulatory bodies, the challenge lies in creating a balanced framework that protects freedom of expression while ensuring accountability for hate-based content. The implementation of the EU Digital services act (DSA) provides an opportunity for N.Macedonia and other Western Balkan countries to align with European standards of digital governance. Policymakers should use this moment to establish national mechanisms for dialogue between state institutions, platforms and civil society. Such mechanisms could include national task forces on online hate speech or annual public hearings where platforms present their transparency data to the public. At the same time, policymakers must invest in education and prevention rather than focusing exclusively on punishment. Legislative reform should be accompanied by national campaigns promoting media literacy, empathy and critical thinking, particularly among young people who are both the most active users of social media and the most vulnerable to its harms.

The media sector plays a central role in shaping public discourse and therefore bears a particular responsibility in countering hate speech. Journalists and editors should apply rigorous ethical standards when reporting on sensitive issues such as ethnicity, gender and migration. Training programs on diversity-sensitive journalism could help professionals recognize how language, framing and imagery can either reinforce or dismantle stereotypes. Media outlets should also adopt internal policies that address hate speech within their own comment sections and online communities, ensuring that moderation is applied consistently and transparently. By setting higher standards of practice, the media can lead by example, demonstrating that public dialogue can be critical without being cruel and inclusive without being politicized.

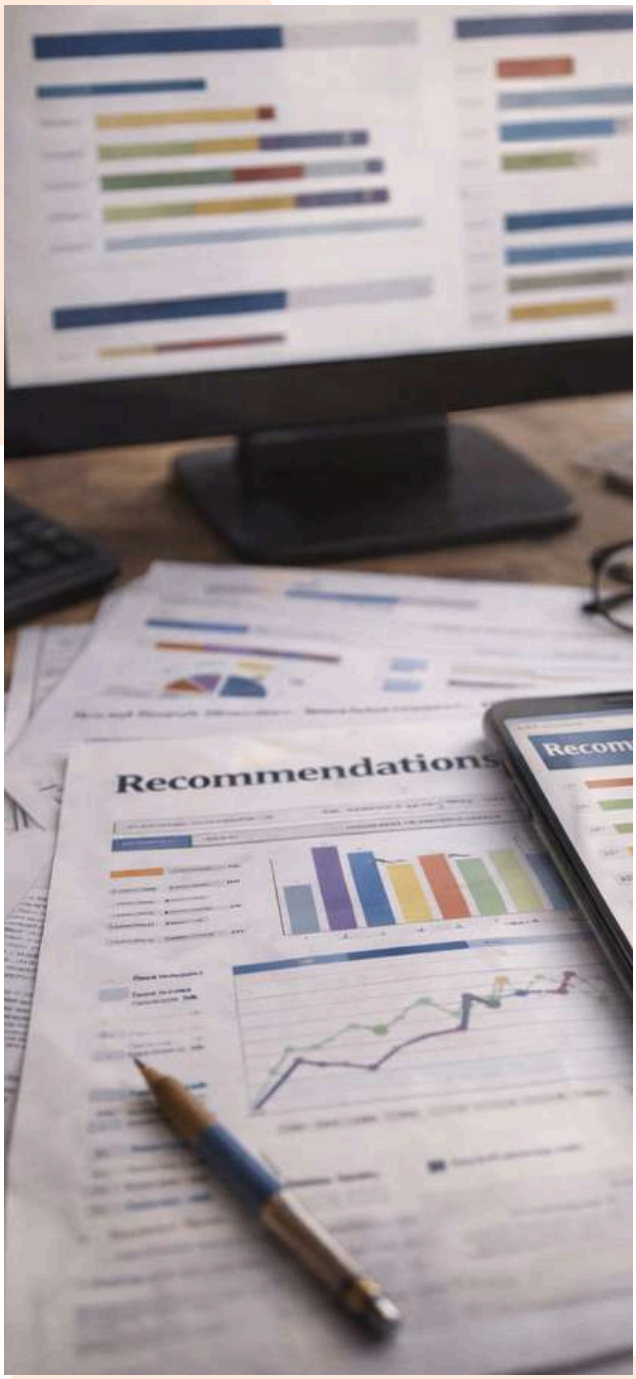
# Recommendations




Civil society organizations, especially those working with youth and marginalized communities, have a vital role to play in building resilience and empowerment from the ground up. Initiatives such as peer education programs, youth-led media campaigns and intercultural dialogue projects can create safe spaces where young people learn to challenge hate through creativity rather than confrontation. These activities should be designed not only as responses to hate speech but as proactive efforts to cultivate digital empathy and civic engagement. Interkultura's experience through the "Words that leave scars" project demonstrates that when young people are equipped with knowledge, tools, and confidence, they can become powerful agents of change producing counter-narratives that shift online culture toward respect and solidarity.

Educators and schools also occupy a critical position within this ecosystem. Integrating media literacy and human rights education into formal curricula can help students recognize manipulation, prejudice and hate-based rhetoric at an early age. Teachers, youth workers and psychologists should be trained to address digital hostility in constructive ways, offering students both technical understanding and emotional resilience. Education can serve as the most sustainable antidote to hate because it shapes not only how people think but how they empathize, engage and communicate in the digital world.

# Recommendations




Finally, these recommendations converge around a broader principle: the fight against online hate speech must be understood as a collective responsibility. No single actor whether a government, a corporation, or an NGO can solve the problem in isolation. Real progress will depend on building trust, sharing data and maintaining continuous dialogue between all stakeholders. What is needed is not simply better moderation but a transformation of the social norms that define acceptable speech and behavior online. This transformation will take time, persistence and creativity, but it begins with recognizing that digital citizenship carries both rights and responsibilities. Words can divide, but they can also heal, educate and empower.

By investing in collaboration, transparency and education, N.Macedonia can take meaningful steps toward creating an online environment that reflects the values of empathy, inclusion and respect. The lessons learned through “Words that leave scars” can serve as a model for the Western Balkans region demonstrating how data, dialogue and community engagement can work together to turn awareness into action and transform digital spaces into platforms for understanding rather than hate.

# Conclusion

The research conducted within the framework of "Words that leave scars" demonstrates that online hate speech in N.Macedonia is not merely a digital phenomenon but a mirror of deeper societal divisions. It reflects long-standing inequalities, political polarization and unresolved questions of identity and belonging that have simply found new expression in the online sphere. The internet, which was once celebrated as a tool of democratization and dialogue, has increasingly become a space where discrimination, fear and exclusion are amplified by algorithms and sustained by silence.

Throughout this study, evidence from multiple sources has converged to tell a consistent story. Quantitative data revealed the scale and typologies of hate speech, showing that gendered and ethnic discrimination remain its most prevalent forms. Meta's transparency reports provided a global perspective on content moderation, highlighting both the reach and the limitations of automated enforcement. Qualitative testimonies from interviews and focus groups added depth to these numbers, offering vivid insights into the emotional and social consequences of digital hostility. Together, these layers of evidence paint a comprehensive portrait of an online environment in which technology and culture interact in complex, often contradictory ways.

The dual reality identified in the research, the coexistence of technological progress and human neglect underscores a fundamental truth: combating hate speech is not only a question of efficiency but of empathy. Automated systems can remove harmful words, yet they cannot repair the trust, dignity, or sense of safety that hate destroys. Algorithms function through logic, while hate operates through emotion. This imbalance explains why, despite improved moderation mechanisms, many marginalized individuals continue to feel unprotected, unheard and unseen. The challenge is therefore not only to regulate speech but to cultivate understanding.

The findings also highlight how political and media ecosystems play an instrumental role in shaping online discourse. Hate speech does not emerge spontaneously, it is often encouraged, normalized, or strategically used by public figures and partisan media to polarize audiences. Each wave of hate reflects a broader atmosphere of uncertainty and competition for legitimacy. Addressing this reality requires both political accountability and media responsibility. Without them, the cycle of provocation and outrage will continue to define public communication, deepening social fragmentation and eroding democratic trust.

# Conclusion

Yet, amidst this troubling reality, the research also reveals grounds for optimism. The very voices that have been marginalized in digital spaces the young, the vulnerable, the discriminated, are also the voices most eager to imagine a different future. In focus groups and workshops, participants expressed not only frustration but also resilience and creativity. Many shared ideas for counter-narratives, campaigns and initiatives that use art, storytelling and humor to promote empathy and inclusion. This willingness to engage, despite the risks of exposure, is itself an act of courage and resistance. It shows that change is not only possible but already beginning to take shape within the very communities that bear the brunt of online hate.

The project "Words that leave scars" is designed to nurture this potential. By connecting research with education, creativity and civic engagement, it seeks to transform knowledge into action. The analytical findings presented in this report form the foundation for the project's next phases: youth workshops, awareness campaigns and public exhibitions. Each of these activities builds upon the evidence collected here, translating data into dialogue and statistics into stories. In doing so, the project embodies the belief that combating hate is not merely a defensive act but a creative one an effort to reimagine how people speak to and about each other in the digital age.

Ultimately, the conclusions of this research extend beyond N.Macedonia. They resonate with a wider regional and global challenge: how to build inclusive societies in an era where communication is instantaneous, borderless and emotionally charged. The Western Balkans, with its history of diversity and division, stands at a critical crossroads where the lessons of the past must inform the ethics of the future. If hate speech represents the scars of history reappearing in digital form, then projects like this one represent the process of collective healing.

The task ahead is neither simple nor quick. It requires collaboration between governments, technology companies, educators and citizens. It demands continuous research, accountability and empathy-driven communication. Above all, it calls for a shared commitment to human dignity as the foundation of all public expression. The internet, after all, reflects what we bring to it. If hate can spread so easily across networks, then so can understanding, solidarity and hope.

The focus group discussions conducted with young participants across several cities in North Macedonia revealed not only the prevalence of hate speech in digital spaces, but also the willingness of young people to challenge it. Many participants expressed interest in developing counter-narratives, promoting digital empathy and contributing to more inclusive online communities. These perspectives highlight the importance of youth engagement as a key component of long-term strategies to address hate speech.

# Conclusion

In that spirit, “Words that leave scars” concludes not as an end but as a beginning a foundation upon which a new culture of digital responsibility can be built. By acknowledging the power of words to harm, we also recognize their power to heal. The future of communication in N.Macedonia and the Western Balkans will depend on how courageously societies choose to make that transformatio

**Funded by the European Union. Views and opinions expressed are however those of the author only and do not necessarily reflect those of the European Union or the granting authority. Neither the European Union nor the granting authority can be held responsible for them.**