

Annual Report on Hate Speech Monitoring 2025

February 2026

digiQ

Summary

The online environment is an integral part of public life. Alongside providing space for creativity, self-expression, and connection, it also enables the dissemination – and often amplification – of hateful narratives present in offline society. Online hate speech has real consequences for users' dignity, mental health, and safety, and cannot be reduced to a purely virtual problem.

During 2025, digiQ's monitoring focused on publicly accessible content on the four most widely used social media platforms in Slovakia – Facebook, Instagram, YouTube, and TikTok – with the aim of evaluating in practice the functioning of mechanisms protecting users from hateful content.

In total, we analysed and reported more than 700 instances of hate speech. More than 80% of them constituted potentially illegal content. The dominant forms were incitement to violence and hatred, including dehumanising narratives and incitement to genocide. Hate speech most frequently targeted LGBTI+ persons, Roma, refugees, and people with a migration background, reflecting the persistent stigmatisation of these groups and the heightened risks they face both online and offline.



The report reveals significant differences in how individual platforms respond to user reporting. The outcome of a submission appears to be largely influenced by the reporting channel used rather than by the severity of the content itself. While specialised reporting channels for expert organisations achieve a high level of effectiveness, the experience of ordinary users is often marked by long delays or a complete absence of response – even in cases involving explicit calls for serious forms of violence.

In the context of the European Digital Services Act, the monitoring results contribute to assessing the practical functioning of user protection mechanisms on social media platforms, identifying their strengths and weaknesses, and providing a basis for further analytical and advocacy steps.

The Online World Is Not Separate from Reality

Digital services have become the infrastructure of public discourse, and online platforms now substitute for public forums, media outlets and community spaces. For a large part of the population, they represent the primary source of information, social contact, discussion of public affairs, and entertainment. Content disseminated in this environment therefore directly influences social norms and the sense of security in society. Dehumanising and hostile narratives — including those promoting violence — affect not only specific victims but also wider audiences who are often exposed to them involuntarily.

Experiences from recent years — including hate-motivated attacks in Slovakia — demonstrate that **the boundary between online and offline violence is becoming increasingly porous**. In 2022, a terrorist attack took place in front of the queer bar Tepláreň in Bratislava, where the attacker shot dead two people and seriously injured another. The perpetrator had been radicalised in online communities associated with militant accelerationism, where he also obtained instructions on how to prepare the attack (Diko & Kőváry Sólmos, 2025). At the beginning of 2025, an attack occurred at a secondary school in Spišská Stará Ves, in Eastern Slovakia, resulting in the death of a deputy headmistress and a female student. The attacker's motivation reportedly included hatred towards women (Hlavanda Gogová, 2025); misogynistic narratives have been strengthening in the online environment for years.

Digital platforms create an environment in which content triggering strong emotions gains greater visibility and spreads rapidly through algorithmic amplification. As a result, platforms play a **significant role in processes of radicalisation and the normalisation of extremist narratives** (Castaño-Pulgarín et al., 2021; Lupu et al., 2023). Monitoring hate speech therefore also has relevance from the perspective of preventing broader societal and security risks.

DigiQ's monitoring fulfils two functions. On the one hand, it has a protective dimension — identified content is reported to platforms so that, where it violates platform rules or the law, it may be removed or restricted. On the other hand, it serves as an **accountability mechanism, enabling assessment of how platforms fulfil their commitments** to users in practice.



What Is Hate Speech?

At digiQ, we rely on the internationally recognised definition of the Council of Europe (ECRI General Policy Recommendation No. 15), which defines hate speech as advocacy, promotion or incitement of denigration, hatred or vilification of a person or group of persons, as well as harassment, insult, negative stereotyping, stigmatisation, threats and the justification of all such forms of expression based on identity characteristics such as race, colour, national or ethnic origin, disability, language, religion, sex, gender, gender identity, sexual orientation and other personal characteristics (Council of Europe, 2015).

Hate speech may, but does not necessarily, constitute a violation of the law.

We focus primarily on illegal speech — i.e., expressions that may fulfil the elements of criminal offences under Slovak criminal law. In addition to illegal expressions, we also monitor, to a limited extent, harmful but lawful speech that does not violate the law but contravenes platforms' community guidelines. This category of content is used primarily to test the effectiveness of platforms' self-regulatory mechanisms.

How We Work with Data

We monitor social media organically and predominantly manually, which allows us to capture context, linguistic nuances and discussion dynamics that might otherwise remain hidden. This approach has its limitations. Given the more limited sample size, the data do not provide a statistically precise picture of online hate speech in Slovakia.

During 2025, we analysed **706 cases of hate speech** across the four most popular social media platforms in Slovakia where publicly accessible content on public pages, channels or communities could be monitored. **Closed profiles, private groups, communities, and channels, as well as private communication, were not monitored.**

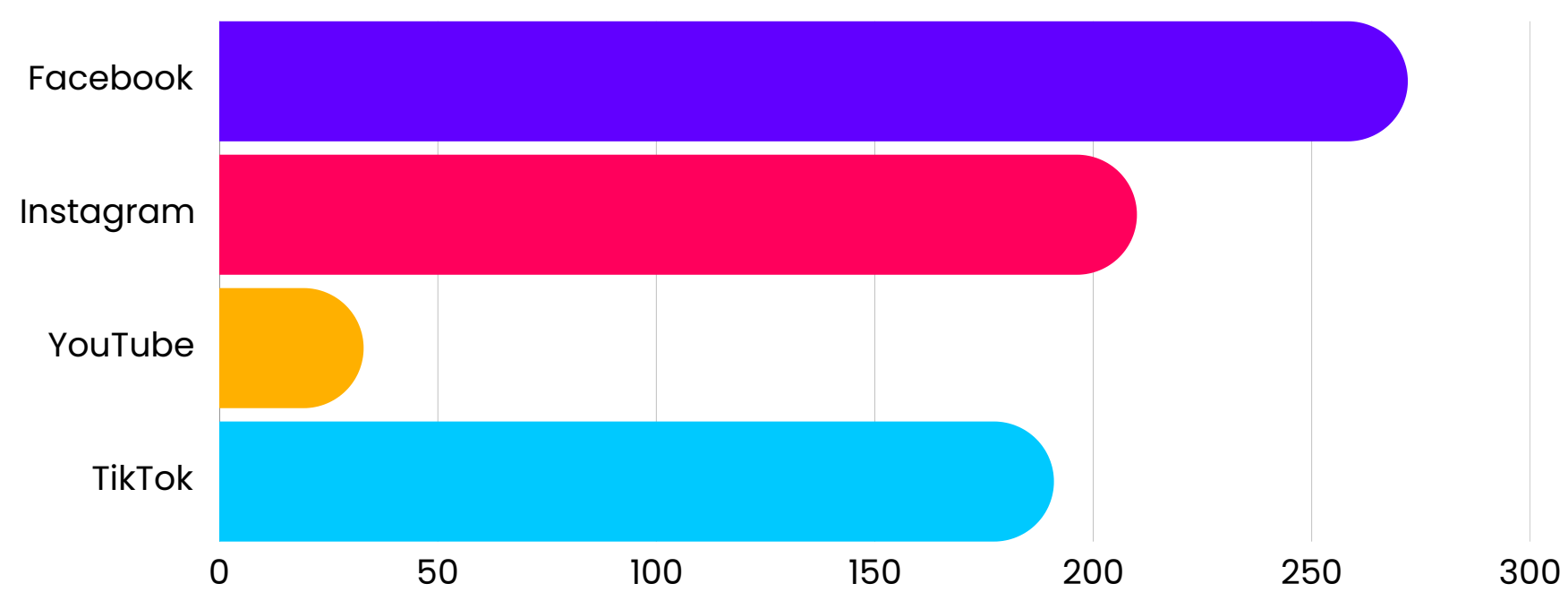
Hate speech was monitored particularly on large and popular channels with significant reach among Slovak audiences. Each instance was assessed on the basis of qualitative criteria to identify hate speech violating community guidelines, as well as potentially illegal expressions under applicable Slovak legislation.

All cases were individually reported to the respective social media platform. Subsequently, the platforms' responses were evaluated. In particular, we monitored the following aspects:

- **Responsiveness:** whether the platform processed the notice and informed the notifier of the action taken;
- **Speed of response:** how quickly the platform assessed the notice and informed the notifier;
- **Action:** whether the platform removed the content or otherwise restricted access to it.

Nature of the Monitoring Data

The distribution of more than 700 reported instances of hate speech across platforms is shown in Graph 1. A total of **83% of the reported content was assessed as potentially illegal**. The remaining 17% consisted of harmful but lawful expressions targeting individuals or groups on the basis of their identity characteristics, which were reported to platforms as violations of their community guidelines.



Graph 1: Distribution of hate speech across platforms in the digiQ dataset

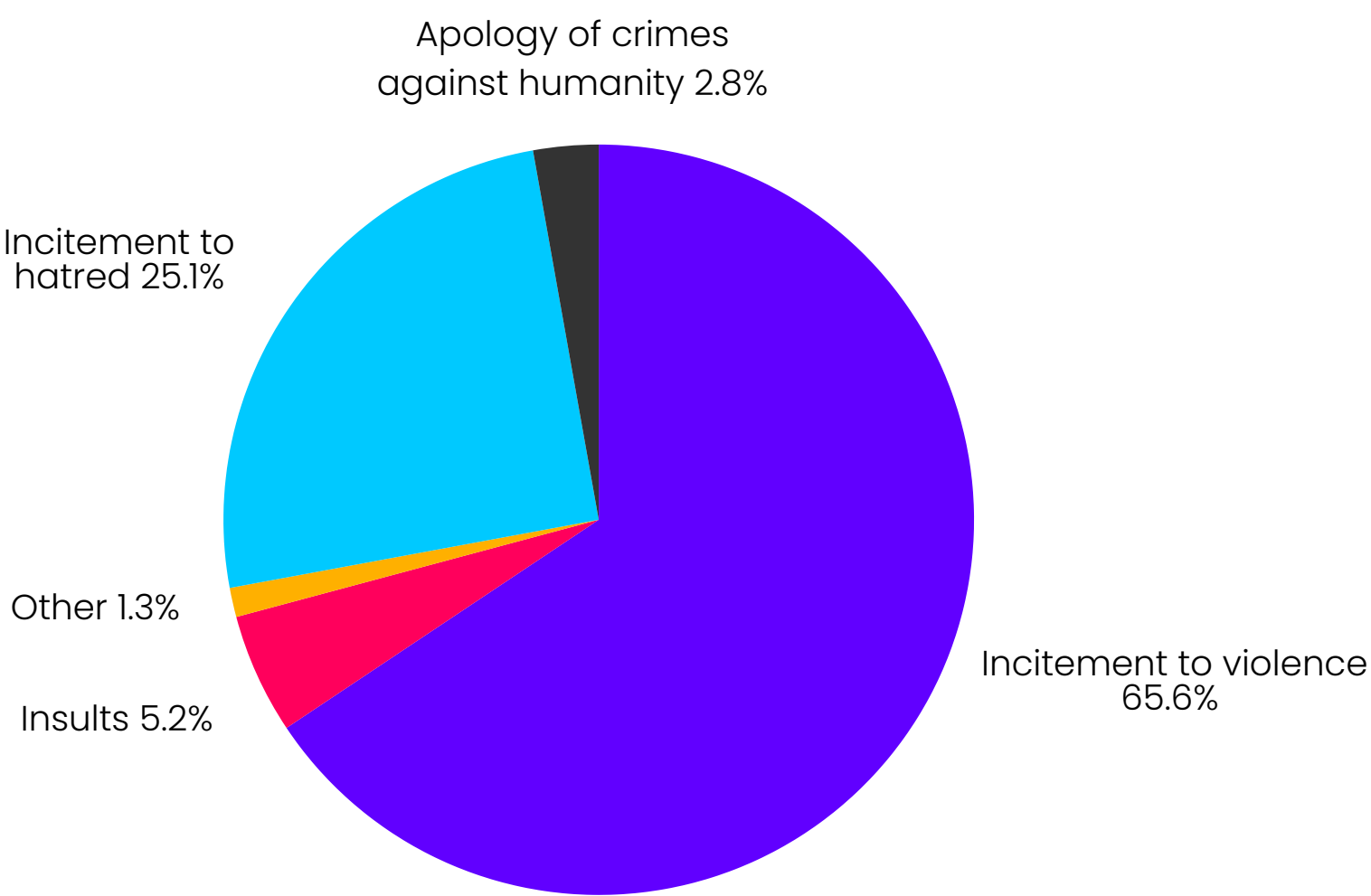
Differences in the volume of identified hate speech do not precisely reflect the proportion of hateful content across platforms, as monitoring was conducted without systematic use of software tools. Nevertheless, these differences may to some extent indicate differing patterns of user behaviour within individual digital environments.

Platform architecture plays an important role – that is, the way discussion spaces and interaction mechanisms are designed. In the case of TikTok, one may hypothesise that short, dynamic, and predominantly visual content reduces opportunities for contextualisation and encourages rapid, emotional reactions, which may include hate speech. In contrast, YouTube’s design appears to support more passive content consumption rather than intensive interaction in comment sections.

In Slovakia, YouTube has nearly four million users according to available estimates – significantly more than TikTok, which still reaches users over 25 to a lesser extent (digiQ, 2025a; DataReportal, 2025). At the same time, the lower number of notices submitted by digiQ on YouTube also reflects the long-term dysfunctionality of YouTube’s reporting mechanisms – since 2023, the platform has not removed a single instance of hate speech we have reported (digiQ, 2025b).

Scope and Forms of Hate

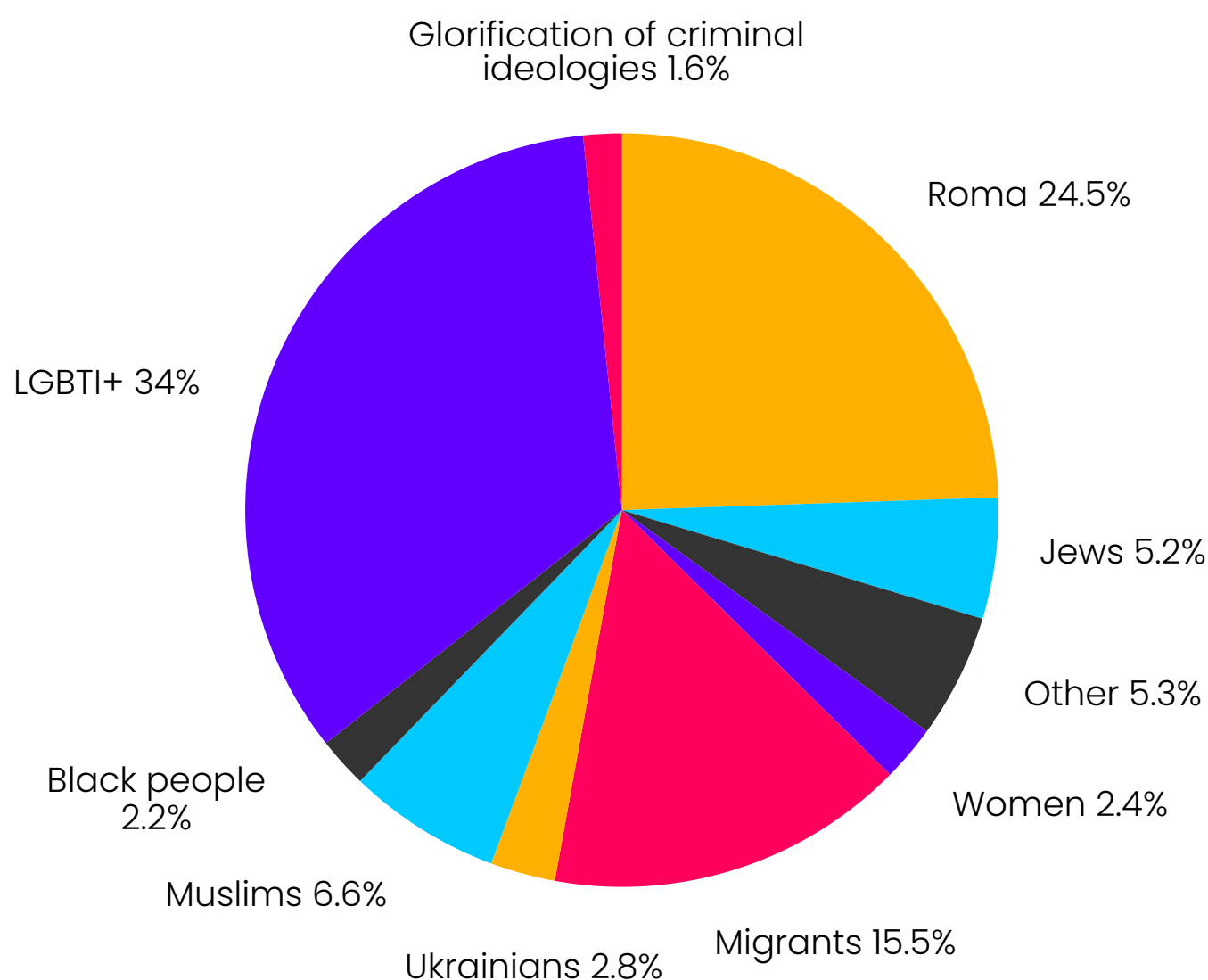
The largest share of the dataset — nearly **two thirds** of cases — consisted of incitement to violence, including genocidal narratives. A further **25.1%** represented incitement to hatred without explicit references to violence. A smaller share, **2.8%**, included narratives justifying or glorifying war crimes and crimes against humanity. Racist, xenophobic or sexist insults reported as harmful content violating community guidelines constituted approximately **5%** of the dataset.



Graph 2: Forms of hate speech in the digiQ dataset

Target Groups and Recurring Patterns

Hate speech primarily targeted groups already exposed to social exclusion, long-term stigmatisation and, in many cases, unfortunate political instrumentalisation offline. The online environment, due to high reach, rapid dissemination and perceived anonymity, often intensifies this pressure.



Graph 3: Target groups of hate speech in Slovakia

How Should Platforms Respond?

Under the Digital Services Act (Regulation 2022/2065), Very Large Online Platforms (VLOPs) – which include all platforms monitored in our study – have concrete obligations when processing notices of hate speech. These obligations apply not only to illegal content but also regulate content that violates community guidelines.

VLOPs must establish **easily accessible and user-friendly notice-and-action mechanisms** (Article 16 DSA), enabling users and trusted entities to notify suspected illegal content, including illegal hate speech. Upon receiving a notice, platforms must act without undue delay, assess the content individually, and adopt appropriate measures (e.g. removal or restriction of access). Decisions must be reasoned and transparent. The notifier must be informed of the outcome, the reasoning, and the means of action (Articles 16 and 17 DSA).

The DSA also requires platforms to **apply their terms and conditions diligently and transparently** (Article 14). VLOPs must ensure that the processing of notices concerning violations of these terms is based on clear rules and that users are informed of the outcome, the reasoning, and the action taken.

As entities with systemic impact, VLOPs are obliged to **assess and mitigate systemic risks**, including the dissemination of hate speech and its impact on fundamental rights (Articles 34 and 35 DSA). This includes ensuring that reporting and moderation mechanisms are not merely formally available but genuinely effective in practice, including in terms of speed and quality of processing. In the Code of Conduct on Countering Illegal Hate Speech Online+, which complements the DSA framework, participating platforms commit to striving to process the majority of reported content within 24 hours of notification (European Union, 2025).

...What Reporting Channels Are Then Available?

Community guidelines represent internal rules created by platforms to protect users from harmful or toxic content. These guidelines are not adapted to local contexts but are often stricter than national laws, prohibiting, for example, a wide range of insulting expressions, harmful stereotypes, and other forms of ridicule.

Thanks to the DSA, users can now independently report illegal content, including illegal hate speech, which they believe violates laws in their country. Article 16 DSA establishes the **Notice and Action mechanism**, under which platforms must provide channels accessible to all users for reporting illegal content. These channels allow notifiers to describe the context and circumstances of the hate speech and indicate which legal provisions have been violated, potentially facilitating more effective processing.

Specialised reporting channels for illegal hate speech are available to trusted flaggers, designated by national Digital Services Coordinators, whose reports must be treated with priority. Within the DSA framework, the status of Monitoring Reporter has also been established, granted by the European Commission to expert organisations participating in annual monitoring exercises to test how VLOPs fulfil their reporting obligations.

Some platforms have proactively created **reporting channels accessible specifically to Monitoring Reporters**, enabling priority reporting similar to trusted flaggers. Among the platforms monitored by digiQ in 2025, such a channel was available on Meta platforms.

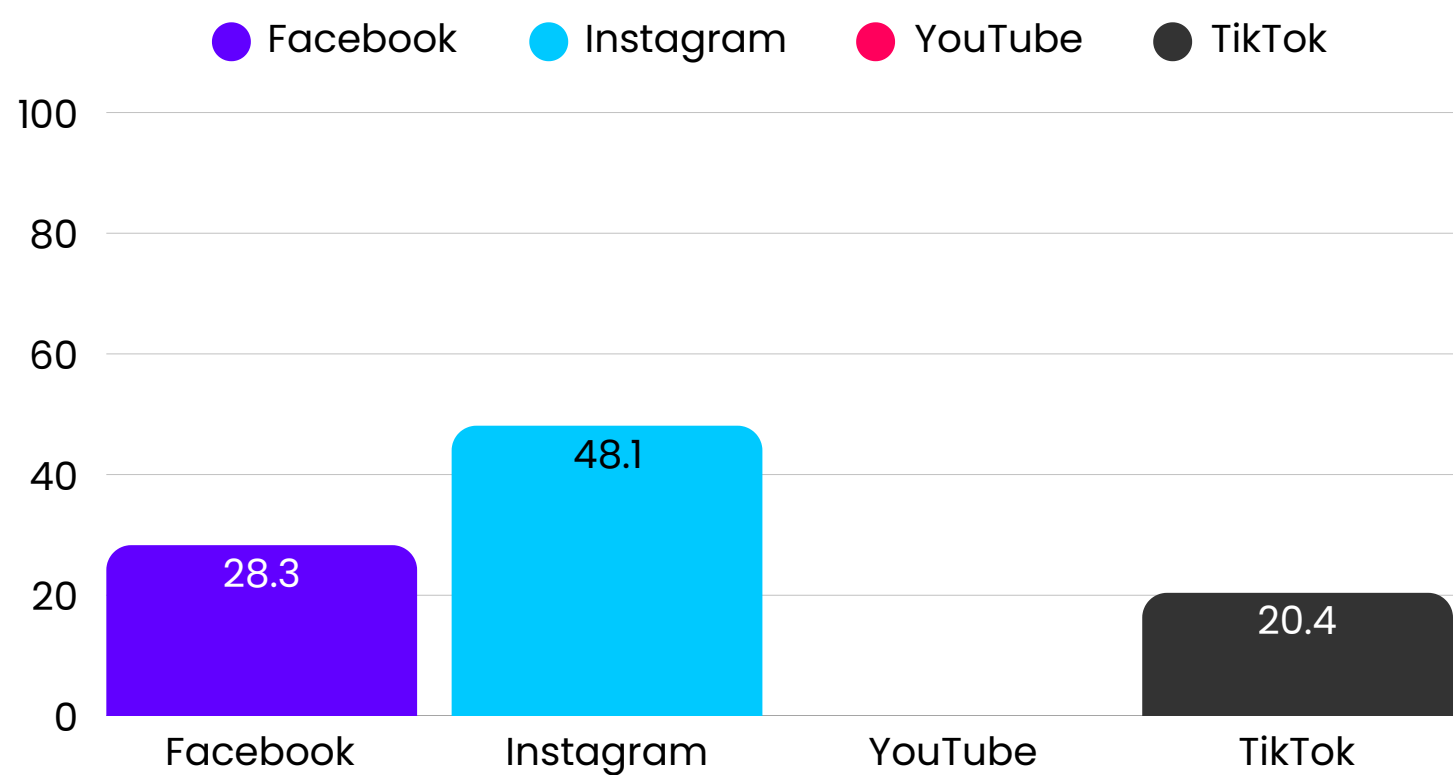


How Platforms Perform in Practice

The DSA establishes clear rules on how reporting should function. But how do the monitored platforms operate in practice? We focused on three key variables: overall **platform responsiveness**, **speed of response to the reported content**, and the **removal rate**. Other forms of action beyond removal were observed only exceptionally during monitoring. These variables provide an empirical picture of how content moderation mechanisms function in the context of obligations arising from the DSA.

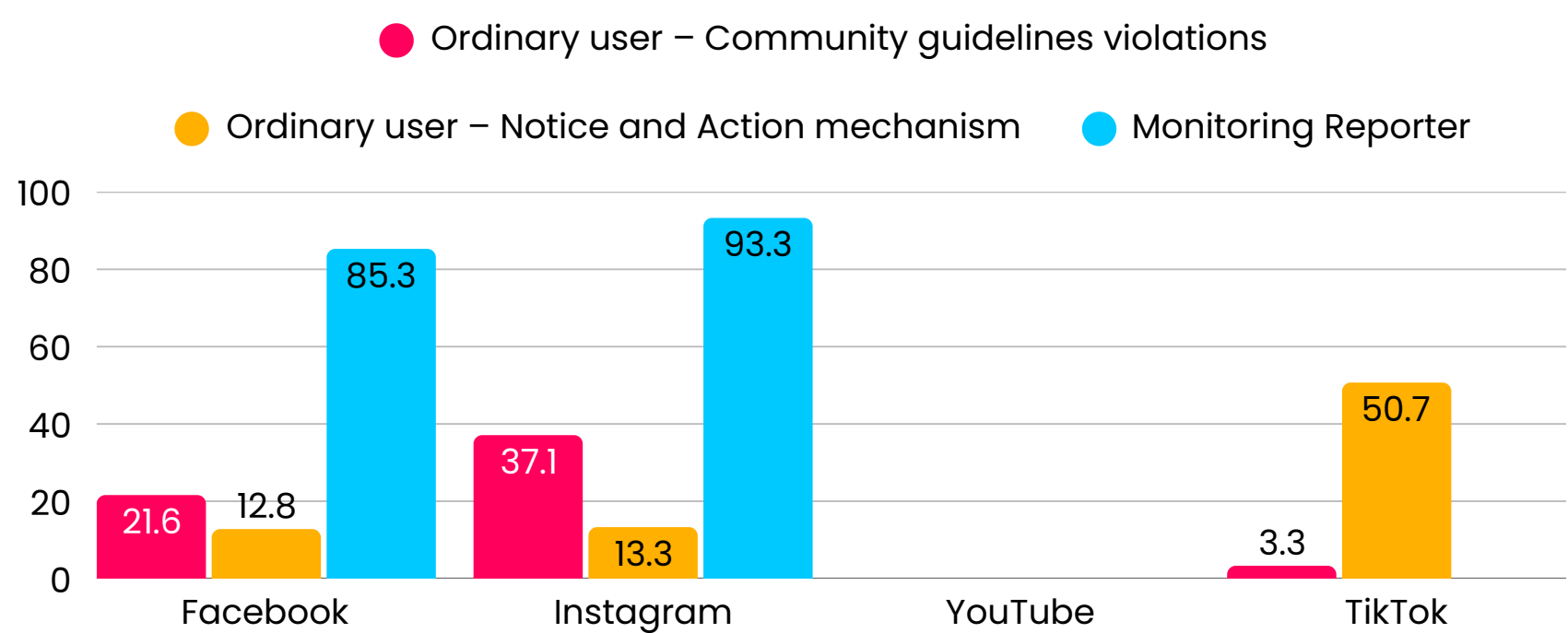
Removal of Content: Effectiveness Dependent on Channel, Not Content

Removal rates varied significantly across platforms. Instagram achieved the highest removal rate, with **48.1%** of all reported content removed. Facebook removed **28.3%** of reported cases – significantly less than Instagram, although both belong to Meta. TikTok achieved an even lower removal rate (**20.4%**), while YouTube did not remove any of the reported cases during the year.



Graph 4a: Overall removal rate of reported content (%)

Differences in outcomes are closely linked to the reporting channels available and how platforms approach them. All monitored platforms allow any user to report violations of community guidelines as well as illegal content under the Notice and Action mechanism. Meta platforms also allow expert organisations holding Monitoring Reporter status to report illegal content through a specialised channel. The following graph shows removal rates through these channels across individual platforms.



Graph 4b: Removal rate by reporting channel.

While Meta platforms provide organisations with Monitoring Reporter status access to a specialised channel, YouTube and TikTok do not offer them.

On **Facebook**, the difference between channels is particularly pronounced. Through the specialised Monitoring Reporter channel, the removal rate reached **85.3%**, whereas under the Notice and Action mechanism for ordinary users it was only **12.8%**. In both cases, comparable illegal content was reported. When reporting violations of community guidelines as an ordinary user, the removal rate was **21.6%**, but a significant proportion of removals occurred only after the content was reported a second time following an initial rejection by the platform.

Instagram also achieved high removal rates via the specialised Monitoring Reporter channel (**93.3%**) and also performed relatively better when reporting violations of community guidelines (**37.1%**). However, under the Notice and Action mechanism, the removal rate for illegal content remained very low at **13.3%**.

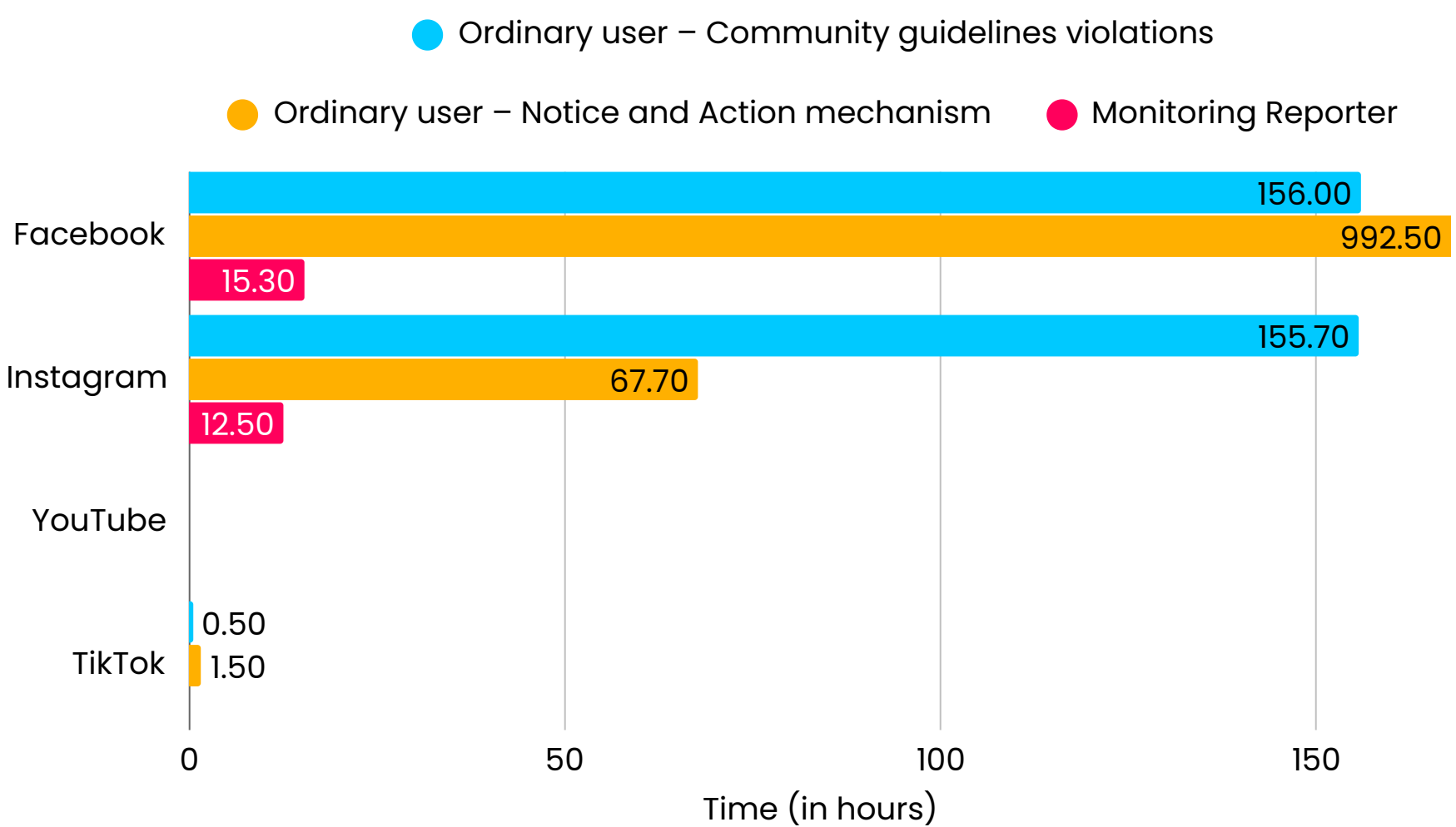
TikTok does not provide a specialised channel for Monitoring Reporters, but the differences between its standard mechanisms are substantial. Under the Notice and Action mechanism, the removal rate reached **50.7%**. When reporting violations of community guidelines, the removal rate was only **3.3%**, and content was often removed only after repeated reporting. Despite rapid processing, this channel resulted in minimal enforcement action against violating content.

A comparison across platforms and channels shows that the **severity of content does not systematically influence removal rates**. When reporting violations of community guidelines, the removal rate for harmful but lawful expressions was approximately **15%**, while for content assessed as potentially illegal, it reached only slightly higher at **20.1%**. Even in cases involving explicit calls for violence, including armed violence, the removal rate did not exceed **21.3%**. In practice, platforms often removed less severe expressions while ignoring reports containing calls to kill or torture.

Speed of Response as a Critical Protection Factor: Processing and Feedback

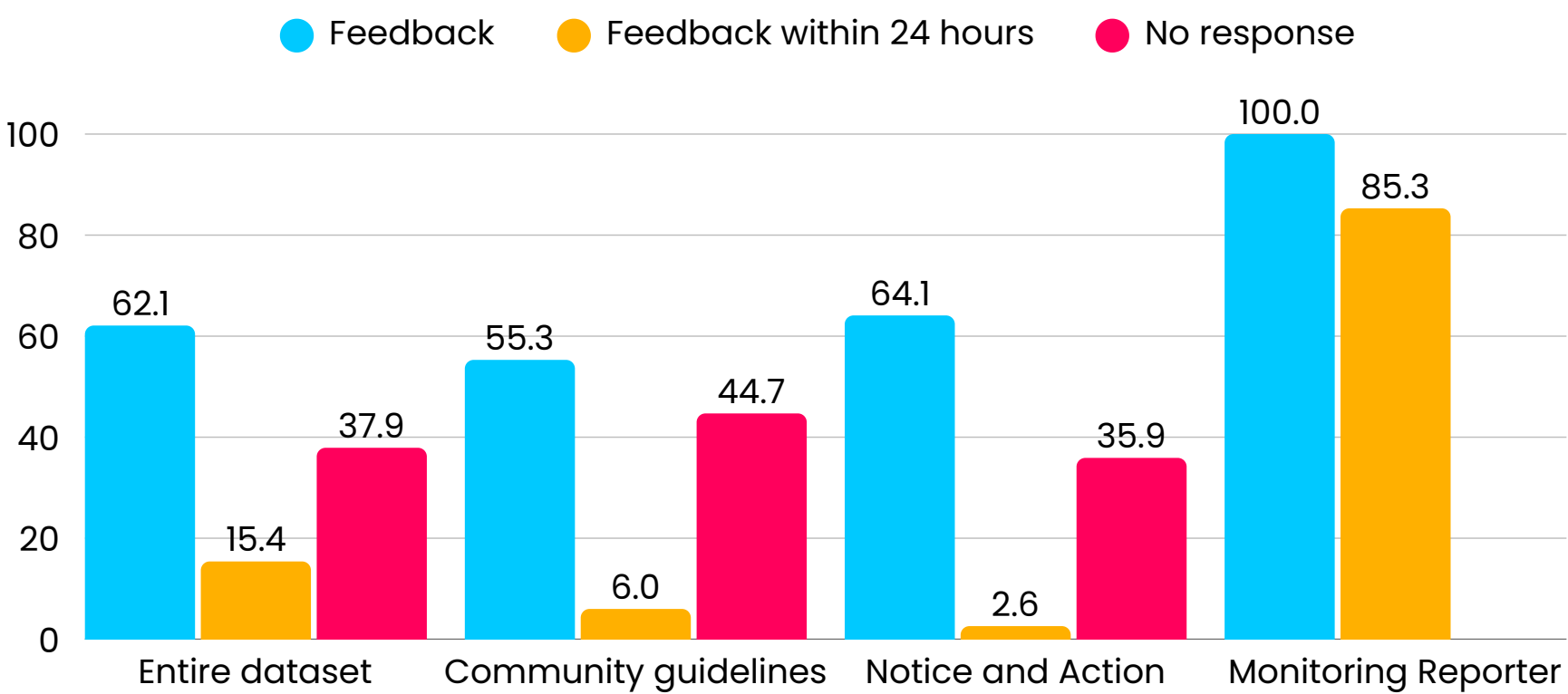
Speed of response was calculated as the time elapsed between submission of a notice and receipt of platform feedback indicating whether the content was removed (or otherwise restricted). The analysis uses median response times. Cases in which the platform did not provide any feedback were included as “no response” and operationalised as ∞.

The results show extreme differences in response speed across platforms and reporting channels.



Graph 5: Median response times by platform. The X-axis scale has a maximum of 168 hours — 7 days. The value for reporting under the Notice and Action mechanism on Facebook does not fit within the graph — the line would in reality be six times longer. Conversely, TikTok’s short response times are barely visible on the scale.

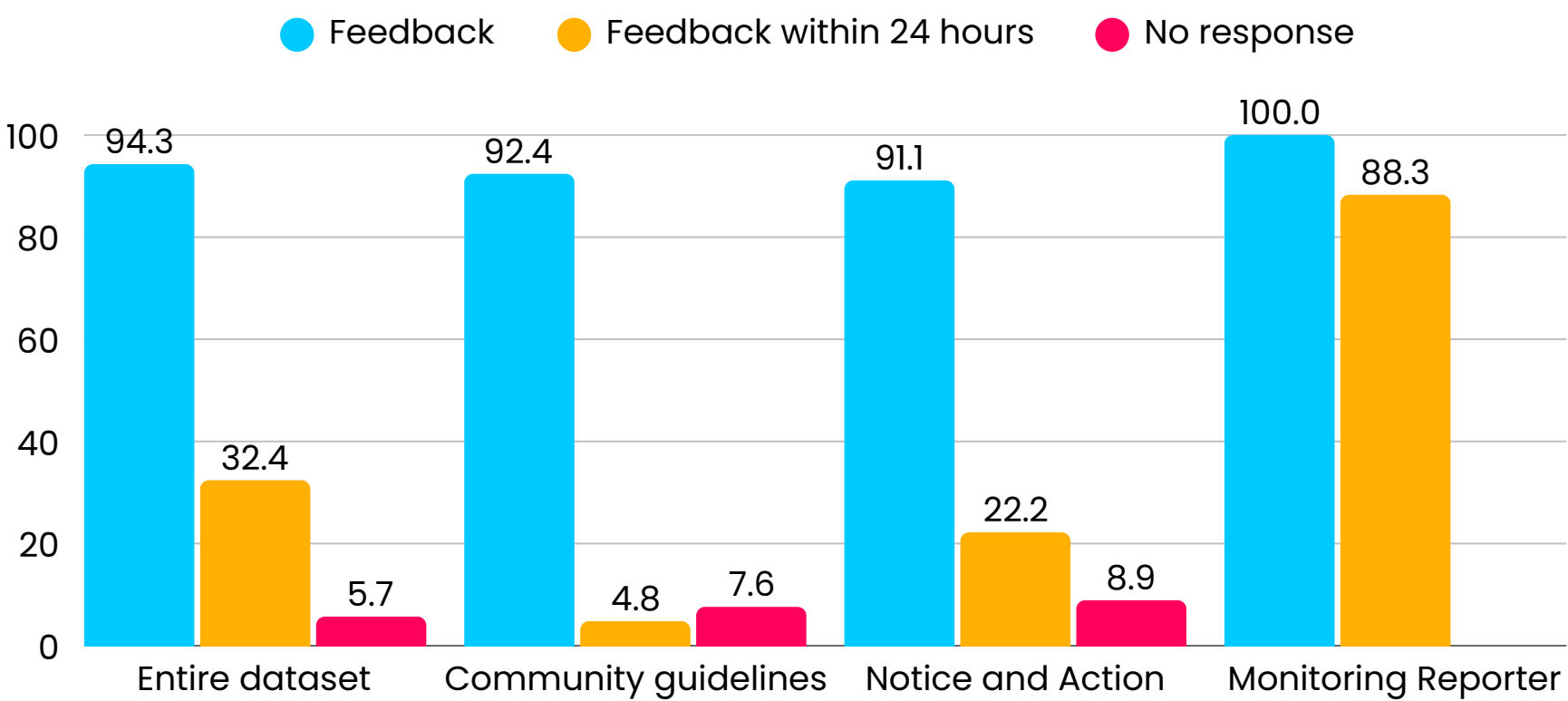
Facebook did not respond at all to **37.9%** of notices. A response within 24 hours — considered the optimal benchmark under the Code of Conduct on Countering Illegal Hate Speech Online — occurred in only approximately **15%** of cases, primarily via the specialised Monitoring Reporter channel, where we observed excellent responsiveness. This was not the case for standard channels, where Facebook failed to respond to at least one third of notices.



Graph 6: Facebook responsiveness (in %)

The median for the Monitoring Reporter channel was **15.3 hours**, whereas when reporting under the Notice and Action mechanism, it reached **41 days**. Even after excluding cases without a response, the average response time (calculated as the mean of response times in cases where any response was provided) remained nearly 23 days, which, from a risk mitigation perspective, renders this mechanism practically ineffective.

Instagram demonstrated significantly better results. Across all reporting channels, the platform processed **more than 90%** of notices. Slower responses were observed under the Notice and Action mechanism and particularly when reporting violations of community guidelines, where **fewer than 5%** of notices were processed within 24 hours. For the specialised channel, we observed 100% responsiveness and predominantly optimal response times.

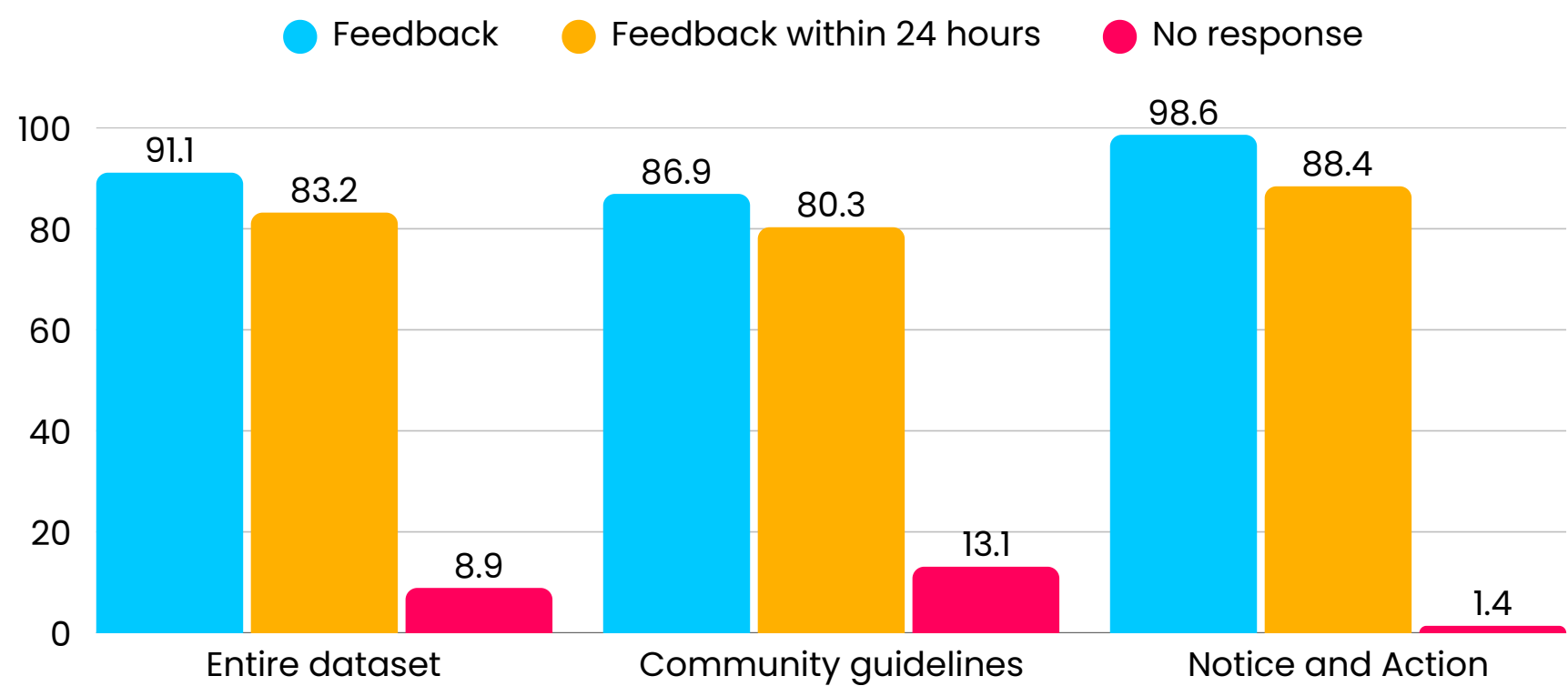


Graph 7: Instagram responsiveness (in %)

The measured median response time when reporting violations of community guidelines on Instagram was nearly one week (**6.5 days**), which – combined with delays and cases without feedback – worsens the typical user experience. When the same user reported content as illegal under the Notice and Action mechanism, the median was reduced by almost four days. Under the specialised Monitoring Reporter channel, Instagram responded with a median of **12.5 hours** and a maximum time shorter than 48 hours.

TikTok presents a contrasting model. The median for the entire dataset was **0.5 hours**, indicating high efficiency. The fact that this exact value accounted for more than half of responses suggests a high degree of automation in processing reports.

We observed very fast processing of notices and high overall responsiveness across both available channels – when reporting content as a violation of community guidelines and when reporting illegal content under the Notice and Action mechanism. In both cases, TikTok responded to **more than 80%** of notices within 24 hours. Even better results were observed when reporting illegal content, where the platform failed to respond in only one case (**1.4%**).



Graph 8: TikTok responsiveness (in %)

However, speed did not translate into proportionate removal rates. This is particularly evident when reporting violations of community guidelines, where the removal rate was extremely low (**3.3%**).

YouTube did not respond to any notices during the monitored period. Despite meetings with platform representatives and repeated attempts to communicate, the enforcement remains practically non-functional for Slovak-language content. This represents a significant systemic risk for Slovak users.

Conclusions and Recommendations

Hate speech in the Slovak online space is not a marginal phenomenon but a stable component of the digital ecosystem. Narrative patterns reflect long-standing social inequalities and stigmatisation, which are not only reproduced online but often escalate into extreme forms. The severity of identified content – including explicit calls for violence – sharply contrasts with the uneven and often insufficient responses by platforms.

The findings demonstrate that protection against hate speech is, in practice, unpredictable and unevenly distributed. On several platforms, access to effective protection depends on the status of the notifier rather than the severity of the violation or the level of risk posed by the content.

Although Instagram demonstrated more positive results than Facebook, both sister platforms showed satisfactory outcomes only when reporting via specialised channels accessible to a narrow circle of expert organisations. Instagram, however, proved relatively more consistent in processing notices. On Facebook, channels intended for ordinary users exhibited extreme delays and a high proportion of cases without response. TikTok demonstrated exceptionally rapid processing, likely due to high automation, but this did not lead to proportionate removal rates even in serious cases. Finally, YouTube proved again to be a platform where reporting mechanisms systematically failed, despite active efforts by digiQ to raise the issue with platform representatives.

The findings show that identical or very similar content is removed depending on who reports it and through which channel. If reporting functions quickly and effectively only within specialised channels, while ordinary users lack access to functional mechanisms, trust in digital services as safe public spaces will be undermined.

Recommendations for Stakeholders

For Online Platforms

- **Increase transparency regarding the effectiveness of reporting channels.** Platforms should provide disaggregated data in their transparency reports on removal rates and processing speed, distinguishing between notices submitted by ordinary users and those submitted by entities with priority status. Such disaggregation is essential for objectively assessing whether reporting mechanisms function for all users, not only for a narrow group of actors. Reporting must remain an effective protection tool, particularly given that engaging in hateful discussions to confront a perpetrator may, due to algorithmic amplification, inadvertently increase the reach of hate speech.

- **Increase the timeliness of content moderation.** Rapid processing of notices is essential for mitigating risks associated with hate speech, including impacts on users' mental health and the radicalisation of vulnerable individuals.
- **Reduce the number of notices without response.** A high share of notices receiving no response is a fundamental problem. Platforms should adjust procedures so that "no response" becomes exceptional rather than a common user experience.
- **Avoid further reductions in the number of human content moderators,** particularly in under-resourced languages, to prevent exclusive reliance on automated processing, which fails to properly account for contextual nuances.

For National Authorities and Institutions

- **Enforce platforms' obligations to notify illegal content to national authorities.** The data show that platforms detect and process vast amounts of illegal content through their own activity and user reports.
- **Ensure stable support and protection for civil society organisations** that possess key expertise on the dissemination of hate speech online, including financial and professional support and protection against targeted attacks aimed at discrediting their work.
- **Systematically work with data and insights from civil society actors** to identify inconsistencies, unequal treatment by Very Large Online Platforms, and broader trends, including comparison with experiences in other European countries. Institutions should use their enforcement powers where civil society lacks the capacity for systematic advocacy.
- **Strengthen prevention** through education, particularly by developing digital literacy and digital citizenship skills in both formal and non-formal education, as long-term tools for preventing the occurrence and impact of hate speech.



Bibliography

Castaño-Pulgarín, S. A., Suárez-Betancur, N., Tilano Vega, L. M., & Herrera López, H. M. (2021). Internet, social media and online hate speech. Systematic review. *Aggression and Violent Behavior*, 58. <https://doi.org/10.1016/j.avb.2021.101608>

DataReportal. (2025). Digital 2025: Slovakia. DataReportal. <https://datareportal.com/reports/digital-2025-slovakia>

digiQ. (2025a). Report on Social Media Use among Slovak Youth. Report. <https://digiq.sk/wp-content/uploads/2025/09/Sprava-z-prieskumu-digitalna-verzia.pdf>

digiQ. (2025b). Results of Social Media Monitoring in Slovakia 2023-2024. Report. <https://digiq.sk/results-of-social-media-monitoring-in-slovakia-2023-2024/>

Diko, L., & Kőváry Sólmos, K. (2025, March). Odsúdený extrémista Slovakbro komunikoval s vrahom zo Zámockej už tri roky pred teroristickým útokom, písali si aj o Teplárni. Investigative Centre of Ján Kuciak. <https://www.icjk.sk/383/Odsudeny-extremista-Slovakbro-komunikoval-s-vrahom-zo-Zamockej-uz-tri-roky-pred-teroristickym-utokom-pisali-si-aj-o-Teplarni>

European Union. (2022). Digital Services Act, 2022, <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32022R2065>

European Union. (2025). Code of conduct on countering illegal hate speech online+, 2025, <https://digital-strategy.ec.europa.eu/en/library/code-conduct-countering-illegal-hate-speech-online>

Hlavanda Gogová, R. (2025, February). Tragický útok na gymnáziu: K zlyhaniu došlo v troch kľúčových bodoch. Učitelia upozorňujú na šírenie mizogýnie. *Hospodárske noviny*. <https://hnonline.sk/tv/relacie/doslova/96194928-tragicky-utok-na-gymnaziu-k-zlyhaniu-doslo-v-troch-klucovych-bodoch-ucitelia-upozoruju-na-siren timerie-mizogynie>

Lupu, Y., Sear, R., Velásquez, N., Leahy, R., Restrepo, N. J., Goldberg, B., & Johnson, N. F. (2023). Offline events and online hate. *PloS one*, 18(1). <https://doi.org/10.1371/journal.pone.0278511>

Council of Europe. (2015). ECRI General Policy Recommendation no. 15 on Combating Hate Speech. <https://rm.coe.int/ecri-general-policy-recommendation-no-15-on-combating-hate-speech-slov/16808b5b08>

About digiQ

The civic association Digitálna inteligencia, known as digiQ, has been actively contributing since 2017 to building a safer and more responsible online environment for all. It focuses on raising awareness of digital citizenship, IT ethics, hate speech, and the protection of human rights in the online environment.

Through non-formal education, campaigns and programmes, digiQ supports the development of digital skills and critical thinking – fundamental pillars of responsible behaviour online. DigiQ's activities include not only preventive initiatives, such as workshops and lectures, but also intervention in response to current challenges in the digital space. As a member of the Digital Coalition in Slovakia, digiQ actively contributes to shaping a better digital future. At the national level, digiQ is a member of the Committee for the Prevention and Elimination of Racism, Xenophobia, Antisemitism and Other Forms of Intolerance of the Slovak Republic (VRAX). At the European level, digiQ contributes to combating hate speech, its prevention, and the promotion of critical thinking online through its membership in the INACH, Zero Hate, and Facing Facts networks. It promotes a responsible approach to the use of the internet and social media in cooperation with the INSAFE network as one of the participating organisations of the Slovak Safer Internet Centre.



www.digiq.sk



digiQ



@digiqsk



digiQ