The Security Risks of Anti-Roma Hate Speech on Social Media Platforms

https://www.e-ir.info/2021/06/12/the-security-risks-of-anti-roma-hate-speech-on-social-media-platforms/

PAVLINA PAVLOVA, JUN 12 2021

Social media platforms have had a transformative impact on how people connect and communicate, how they express, search for, and encounter information, and how they organise. Many changes have been positive but the enabling dynamics of these platforms have also opened the possibility of being exploited by organised hate groups, for organising attacks or intimidating and harassing members of minorities. The historical discrimination toward various ethnic minorities has found new channels and flourished online. Already back in 2003, the OSCE Ministerial Council Decision on Tolerance and Non-discrimination had acknowledged that racist and other hateful content on the Internet can instigate hate crimes. A recent statement by the UN Special Rapporteur on minorities issues highlighted the persistence of these issues, when accusing the propagation of hate speech through social media of contributing directly to the rise of hate crimes against minorities, and calling for this “poisoning of minds” online to be acknowledged and confronted.

Hate online content is not an anomaly, but a permanent feature of the internet that disproportionately affects vulnerable people. With an estimated 10 to 12 million Roma people living in Europe, these communities are the continent’s biggest ethnic minority. At the same time, 80% of Roma men and women are living at risk of poverty. Their low economic and social status continues against a backdrop of persistent discrimination — they are the most disliked and discriminated minority in Europe, with more than half of citizens having a negative view of Roma. Such attitudes are amplified in online spaces, sometimes artificially and in an organized manner. There has been a rising wave of anti-Roma rhetoric in recent years, and the coronavirus pandemic has further reinforced this trend. Despite the urgency of these issues, the impact of social media-fuelled hate speech on minority and marginalised groups has not been fully acknowledged by the providers.

Hate speech turning into hate crimes

Online hate plays a significant role in normalising racist and xenophobic attitudes. Posts and comments depicting negative stereotypes and prejudices can increase interethnic tensions. They can also lead to a negative spiral when manifestations of hate foster more hate or even escalate into physical harm.

Racist incidents usually develop according to the following paradigm: a right-wing group publishes a xenophobic post about the Roma. That post provokes a growing level of hostility online. For example, others voice their support for these views in the comments section, the post becomes known, even viral, and then, there is an attack. This information is not verified and is often created specifically to justify the attack. In 2018, there were attacks on temporary Roma settlements in Ukraine, which were preceded by overtly anti-Roma posts that called for violence. Such and similar cases demonstrate that decisions on online content have a major role in shaping events with life and death consequences.

The pandemic has brought a new wave of anti-Roma sentiment and hate speech, fuelling violence in several European countries, where the Roma are persistently blamed for violations of the quarantine restrictions, as well as spreading or causing COVID-19. One of the distinctive features of such posts is that they provoke an intense response from other users who engage in hate-speech comments inciting hatred, violence, expulsion, bullying and

perpetuating social stigma. Another trend is the use of derogatory metaphors when referring to Roma, which serve the very purpose of dehumanizing the Roma as a group. This often happens in online groups, which can be created for the specific purpose of spreading hate against Roma.

Particularly dangerous are comments that speak about the incompetence of the police in sustaining public order and preventing investigating alleged antisocial behaviour or crimes. Such reactions incite people to take "matters into their own hands" with a claim that they are the only ones who can restore the social order. By doing so, the hate groups delegitimize state institutions and create a justification for violent actions.

Content moderation practices

Major providers of digital services have undertaken steps toward more responsible content moderation. Yet the efforts are mostly reactionary – after negligence to online hate speech resulted in severe real-life consequences. Providers are also often overly US-centric, and the marginalised communities do not receive adequate attention. But with growing possibilities of spreading hate speech on these platforms, more active and targeted efforts are needed to prevent future harm.

Efficient content moderation is not an easy task. The providers are left calibrating the red lines of what should be considered as content that can stay and content that needs to be removed, and the risk of overregulation is substantial. Freedom of expression can be legally limited in cases where it is inciting violence. In such cases, there must be a close nexus between the expression in question and a substantial risk of harm. However, despite many cases providing such a nexus, hateful content remains present on social media platforms.

The providers are inconsistent about enforcing their existing standards, which have for long prohibited bullying, hate speech, and incitements to hatred and violence. They also often lack the human content moderators that would understand the context of such situations. Moreover, it is important to consider that such binary decisions, while being an important human rights issue, are not the only way how to lower the security risks. Social media algorithms can also regulate the overall flow of posts. The decisions about what platforms choose to amplify or add friction to are the most important in content moderation, and the most vital in high-risk situations.

As Facebook has shown in emergency situations in the past, it can effectively limit the spread of harmful content that its systems predict is likely to violate the company's community standards. The company can reduce toxic content by turning the dial down on spreading inflammatory posts, limiting its distribution and engagement. This method discourages borderline content and makes the platform a safer place. These methods are used temporarily in high-risk locations. Recent examples include risk related to civil unrest and insurgency episodes in Myanmar, Ethiopia, and Sri Lanka. The security risk is then lowered not by changing the rules on hate speech, but due to reduced virality of problematic content in personalized news feeds. This method addresses the main difference between online and offline hate speech—the possibility of rapid spread—when the hate statements are shared, encouraged, repeated, and engaged in the comments section leading to harm and crimes in real life.

Further automation and content-recognition technologies can be primarily helpful in cases of ethnic tensions with a high potential of escalation into a conflict. What remains problematic is that content is often deeply context-dependent, which means that automation can fail or generate false positives, and lead to direct or indirect violation of users' rights to freedom of expression. Automation still falls short on understanding and interpreting the context linked to the post, the user's intention as well as the linguistic and sociological context of the post. Evaluation of hate speech is highly contextual. This presents an intractable problem for global platforms searching for scalable solutions. Hate speech is often not clear, it tries to play with words, forms, and rules to escape the formalised language. On the other hand, if there is no guarantee of a qualified human review, it can restrict posts that are legitimate and lead to larger collateral damage, especially as the rules of the enforcement lack transparency.

These approaches are unlikely to work in isolation and will need to be designed to work together. Moreover, responsibility for preventing hate crimes is on the side of states as well. While legal rules and norms on what constitutes hate speech vary around the world, content that calls for riots or crime can qualify as an intentional act to

incite racial hatred and is illegal in many countries. However, what is missing are precise legal definitions and mechanisms for the effective monitoring of their implementation, which coupled with victims low level of trust towards the police, leads to a high level of under-reporting of such cases. What is also missing is specialisation and capacity on the ground, in criminal justice systems—among law-enforcement officials and prosecutors. The authorities should invest in building their capacities and trust to protect individuals and communities from harm, including by referring to the experience of affected people and civil society organisations that have been documenting and analysing hatred online translated into real-life incidents.

Besides the challenges posed by legislation and its applicability, there are also practical difficulties of removing hate-inciting material. Once the content is posted on the Internet, it is shared on several accounts and platforms making it problematic to remove all the content completely. The collaboration between the companies and governments remains for now limited, split over questions about effective approaches to moderating content and the appropriate regulation.

The way forward

Current content moderation methods are unsustainable. The standards need to be considered in terms of the communities they affect, particularly in cases of marginalized and vulnerable communities in high-risk environments. Both platforms and states should guarantee meaningful participation of targeted people and their perspectives in the development and consultation around the creation of standards and the way they are enforced.

To support this process, we need more transparent processes and more efficient data collection. Disaggregated data is necessary to understand the extent and impact of persistent hate speech on the communities. The platforms that influence globally must start moderating with a local context in mind. They need to take proactive measures and add much-needed context and resources to the way they operate, monitor and enforce their standards. Adequate content moderation must facilitate a more security-sensitive and risk-responsive online environment.

Content on social media leading to offline harm is not confined to Roma in Europe. It is a global problem. While the people most suffering because of hate speech are those who are targeted, hate speech divides societies, and increases the overall degree of aggression and insecurity for all. Less hate speech online means more security and more resilient societies. Platforms should not wait until the harm is done to act on it.

About the author:

Pavlina Pavlova is a former official and currently a consultant at the Organisation for Security and Cooperation in Europe (OSCE) working on digital security assessment and capacity building of Roma human rights defenders in the OSCE participating States. She has worked with hate crime victims and gained first-hand experience about the risks and harm that inadequate content moderation, related legislation and the lack of enforcement possess for Roma in eastern Europe.