Study on

the scope of hatred in Estonian social media

conducted during September 2022

Authors Kelly Grossthal and Mari-Liis Vähi

Introduction

Estonian Human Rights Centre is the only organization in Estonia that actively monitors manifestations of hatred and raises awareness of its dangers. The Estonian Human Rights Centre is a partner of the European Commission for monitoring of hatred incitement tendencies on all social media platforms. The legal framework on hate speech falling within criminal law is very limited and criminal action is rarely taken. Therefore, it is both challenging and difficult to combat hatred in Estonia. In addition, fake news being spread in recent years have contributed to an increase in hate speech across social media. This study is aimed to fill this gap and determine the scope and narratives of hatred found in Estonian social media. The study was conducted within a project "Examine. Participate. Change" aiming to improve resilience and social cohesion and integration by strengthening fact-based media, media literacy and civic engagement and cooperation in the field of education, training and exchanges in Estonia. The project was financed by the Embassy of the Federal Republic of Germany in Estonia.

Background and needs analysis

In common language, "hate speech" refers to offensive discourse targeting a group or an individual based on inherent characteristics (such as ethnicity, religion, sexual orientation, age, disability or gender) and that may threaten social peace. Hate speech promotes hatred or violence against those groups or group members. Hate speech is not protected under freedom of expression as it is an abuse of freedom of expression. We are free to express ourselves, even to the extent that our opinion may offend, shock or disturb others. But not everything is acceptable as free speech. When people start publicly inciting to violence, hostility or discrimination against certain groups, then this is hate speech not free speech.

This also means that the state can legally prevent hate speech and punish for such expressions. The purpose of hate speech is to humiliate and degrade a person because of who they are, not because of what they have done. Hate speech was used as a tool in, for example, the Nazi holocaust and in the genocide in Rwanda.

Hate speech is always contextual, it depends on the environment, the speaker and the target group. To define hate speech, various components must be considered, including the content of the expression, the (written or oral) tone, the (individual and collective) targets, and the possible consequences or effects. It can be challenging to tackle hate speech as it refers to a variety of speech acts and other ill-behaviors, ranging from penal criminal acts to speech that is uncivil and disturbing, and yet tolerated.

Next to recognising and understanding phenomena, the legal landscape plays an important role in addressing and curbing hate speech. Estonian anti-discrimination legislation is based on §12 of the Constitution, which prohibits discrimination and incitement to ethnic, racial, religious or political hatred, violence or discrimination is prohibited and punishable by law. The Penal Code includes provisions which prohibit incitement of hatred as well as breach of equality in general. The provision prohibiting incitement of hatred is as follows: § 151 (1) Incitement of hatred. Activities which publicly incite to hatred, violence or discrimination on the basis of nationality, race, color, sex, language, origin, religion, sexual orientation, political opinion, or financial or social status if this results in danger to the life, health or property of a person is punishable by a fine of up to three hundred fine units or by detention.

Incitement to hatred or violence is a legal term which is quite often used synonymously with hate speech. At the same time, hate speech has a wider scope compared to incitement to hatred or violence since hate speech covers broad forms of expressions which advocate, incite, promote or justify hatred, violence and discrimination against a person or group of persons for a variety of reasons.

In practice, use of this prohibition is limited and it has been applied on only a few occasions, for example in 2019 and 2020 there were zero cases reported under the § 151 (information about 2021 is not disclosed yet). The problem lies in the wording of the provision, according to which only such incitement of hatred is punishable, which poses an immediate danger to life, health or property of a person. Hence, law enforcement have limited possibilities to hold hate speech perpetrators accountable in Estonia.

Although hate speech can take place both offline and online, the latter is recognised as a growing online issue which can negatively impact a person's mental health, general wellbeing and online engagement. It can also, in the most extreme cases, lead to harassment and violence offline. As it is known, the past decades have been marked by the expansion of internet and social media platforms. The rise of online hate speech has been accompanying this growth, leading national and European institutions to create policies in order to tackle this phenomenon. Arguably, the worldwide pandemic of covid-19 has increased the time spent on the internet and an important resurgence of online hate speech has been noticed. The lack of systematic moderation by the IT companies

is making the internet become a space encouraging the rise of extremists and illegal hate speech. Although the European Commission and civil society organizations have had different initiatives to tackle online hate speech, the need to monitor hate speech and illegal content on social media is an ongoing necessity.

Methodology of social media monitoring

During a month-long period, two social media monitoring sessions were conducted to examine social media posts created both by Russian-speaking residents of Estonia and their Estonian-speaking peers. 51 posts with hateful content were found in Estonian and 52 posts with hate speech were found in Russian. The study compares the content of those posts made in Estonian and Russian and establishes the substantial narrative differences. For each post, social media platforms and three narrative aspects were marked down.

The first narrative question aimed to determine the protected group the post targeted. In order to have more comparable factors to analyze, two protected groups were chosen. The first group was gender and sexual orientation, and the other was nationality, ethnicity and race. These two groups were chosen because these are the two categories of groups that have been the regular and systematic target of hate speech in Estonian society in recent years (according to the previous monitoring sessions the centre has conducted).

The second narrative question was aimed to determine the type of hatred used in the post. It was determined whether the hatred was a direct call to action, incitement of hate, statements of inferiority, expressions of contempt, disgust or dismissal, cursing and calls for exclusion or segregation. The third narrative question determined whether the post used a direct language or a metaphor. During the monitoring, tendencies and patterns (such as specific words used) of hatred found in social media were identified and marked down.

For each social media post consisting of hate speech, the posts were reported and it was also measured how quickly, if at all, social media platforms removed reported posts that contain hate speech, and whether this is affected by the language of the post. For the monitoring four mainstream social media sites were chosen, i.e. Facebook, Instagram, TikTok, Twitter. For future social media hate speech monitoring in Russian in Estonia, Telegram would be also a mainstream platform that could be monitored, however, for comparison purposes forementioned platforms were chosen.

Results of the study

In general, the monitoring staff noted noted that there is a lot of hateful and hostile content. However not all of it would be considered hate speech or incitement to hate that is against the social media platforms' community standards or rules of conduct (and against local laws). In both languages, it was noticed that people have learned to spread hatred by overcoming social media hate detection AI models - for example, using metaphors and hidden threats. As well, monitoring staff noticed that people warned each other from reporting and getting blocked. During the monitoring of social media, platforms, groups and pages that consist of more hateful content were identified. In Russian, hatred was monitored from LGBT+ groups on Facebook groups in Russian. As there is less news on the LGBT+ topic in the media right now, some of the reported comments were posted a few years ago.

In Estonian, hate speech was found in comments from Facebook pages like Delfi, Postimees, Objektiiv, Eestinen, Uued Uudised and under TikTok videos like comments under Estonian non-white, transgender bloggers' videos. Unlike posts of public figures, under mass media news pages like Delfi and Postimees, there were a lot of comments that were hateful, probably because of the massive load of comments and the little resources to monitor, report and get them removed. Individuals seemed to keep their accounts clear and monitor comments under their posts or even disable commenting under the posts. It was also noticed that younger people in Estonia mostly tend to express themselves more in TikTok and the older generation prefers Facebook for that. Comments in TikTok and Instagram tended to be more short and simple than the ones on Facebook or Twitter.

Narratives of hate speech determined during the social media monitoring

Throughout the entire monitoring session, 103 instances of hate speech were detected. As previously mentioned, Facebook was the most popular tool for spreading hateful messages among the five big social media platforms and the findings confirms it, as; 87 of the found content was posted on Facebook, 12 on Tiktok, 2 on both Instagram and Twitter.

Majority (i.e. 80 comments) of these messages were directed against gender and sexual orientation (predominantly against the latter), while the remaining (i.e. 23 comments) were against groups on the basis of their nationality, ethnicity, and race. For both groups name calling was apparent in almost all of the comments. For the first group, the groups of people on the basis of their sexual orientation were frequently targeted. In both languages, non-heterosexual people were often compared to pedophiles, zoophiles or apes,

or that they are carrying a disease. In Estonian, word “pede” (in English “fag”) was the most common word used in comments against groups of people on the basis of their sexual orientation. The word “pede” was used in 9 comments reported. In Russian, there were more comments with directly threatening language used against non-heterosexual people (e.g. inciting to “burn them”), as compared to the content in Estonian, where the language used was mostly statements of inferiority or expression of contempt, disgust or dismissal. Few times there were also hateful comments against transgender people referring them as “it” or saying that, for example, “trans woman is not a woman”.

In the comments against people on the basis of their nationality, ethnicity or race, black people were the most frequent target of hateful comments (i.e. 17 comments). The word “neeger” (in English “nigger”) was used four times. In comments against people of African descent, people called for colonization, slavery or segregation. Narratives that came out of the monitoring were that people of African descent are lazy “unless under white power” or that people of African descent should “go home”, we should send them “back to Africa” or that they “don’t belong here”. There were four hateful comments against Russians, one against Ukrainians and one against the Arab people.

The majority of messages included the elements of expressing contempt, disgust, or dismissal; however, they also noticed that there were messages including calls for exclusion or segregation, statements of inferiority, incitement of hate, and direct calls to action (including incitement to violence). However, some comments also used a language that, for example, used both statements of inferiority and expressed contempt, disgust or dismissal. It was noted that 93 of these instances used a language that directly targeted the victim group while the remaining (10 comments) resorted to metaphors for delivering their hateful message. From direct hate speech messages the most frequent ones were calls to “burn yourself”.

Entirety of the content in the Russian language was written in a direct manner. On the other hand, 41 (out of 51) Estonian messages were written in a direct manner, while 10 of them resorted to metaphors. In contrast to hate speech against LGBT+ community in Estonian, in Russian there were calls to bully LGBT+ kids. In Estonian there were words like “pede” and “neeger” frequently used. In Russian there were also a lot of similar words used against LGBT+ community. Words that were used against protected groups of people on the basis of their sexual orientation were “голубой”, “петушара”, “пидар”, “педики”, “пидарюги”, “гомосек”. Words that were frequently used against protected groups of people on the basis of their gender were “извращенец”, “мерзость”, “твари”, “гнойные”, “педофилы”.

Social media response to the reporting

The monitoring staff reported the content that consisted of hate speech or incitement to hatred found on the social platforms was immediately reported to the appropriate departments. Reports included information about the post's contents and visual evidence, and they were directly addressed to the community guidelines that this content was violating. Out of 103 reported posts, 38 of reported posts (26 for Facebook, 12 for TikTok) were taken down on the first day, where 28 were Estonian while the remaining 10 were Russian content. On the first day, Facebook did not find it to be against their community standards for 30 posts and decided not to take them down. On the second day, Facebook decided not to take down another post. On the second day, Instagram took down one of the reported posts and, on the third day, another one of the two reported posts on Instagram. On the third day Facebook took down two more posts and decided for four posts that it is not against their community standards and thus, they will not remove them. After a week from reporting, there were no more posts taken down. Therefore, 42 posts were taken down out of 103 reported posts. This shows that the social media platforms removed only 40.78% of the reported posts.

Twitter did not take any action to the reported content within the monitoring period (1 week). Instagram removed both of the posts (on the second and third day). TikTok removed all of the reported posts (12) on the first day. Facebook took down 29 out of 87 reported posts (with a ratio of 18 Estonian and 11 Russian). This means Facebook has only taken action against 33% of the reported posts. The monitoring staff was also surprised to see that Facebook did not take down some of the content that was clearly hate speech and written in a direct manner but it did take some of the comments that were less direct and hateful (but still against their community standards) - therefore monitoring staff found the Facebook monitoring system to be unclear and unsystematic reacting to hate speech. There were a significant number (20) of reported comments that did not receive any reaction from Facebook's support staff. There were no significant differences in Facebook content moderations in Estonian and in Russian.

Recommendations

- **Current study confirmed the centre's previous monitoring exercises findings that spread and use of hate speech is a persistent problem in Estonia and especially in the online sphere. Therefore there is an urgent need for a national action plan on combating hate speech involving intersectional actors.**

- There is a need for more public discussion about the notion of hate speech and freedom of speech. There are many misconceptions among the general public as a result of misunderstandings and disinformation. The state should initiate or fund programs to raise public awareness on the effect of hate speech on the targeted groups and society at large.

- As the majority of reported hateful messages were directed against gender and sexual orientation (predominantly against the latter), children and adolescents should therefore be educated in school about gender diversity, including lesbian, gay, bisexual and transgender (LGBT) issues. Psychological and sociological research signals that heteronormativity, homosexuality non-acceptance, and negative attitudes toward LGBT community are associated with lower levels of education and intelligence. For students who show gender nonconforming behavior, schools can create a safe climate if they succeed in reducing homophobic and transphobic discrimination.

- Due to the fact that some political parties are involved in spreading hateful, discriminatory messages, it is important to encourage the parties to adopt the code of ethics and have the instruments to react if the politicians break it.

- There is lack of awareness and unwillingness to address the issue of hate speech that in turn contributes to a climate of impunity for abusers. Therefore, human rights education in the field and general advocacy work is very much needed.

- In Estonia, the hate speech laws are too soft and the current Penal Code is futile against hate speech for its wording since it requires words to be accompanied by direct danger to one's life or well-being. The Penal Code has to be amended so that the state could also react to instigation of hate and calls to violence.

- Proper research is needed to deduce the causes, prevalence and impacts of hate speech in Estonia. There is a need to further research (on a larger scale and regularly) the sentiments and narratives spread in Estonian social media in both Estonian and Russian in order to effectively tackle language and culture specific hate speech.

ESTONIAN HUMAN RIGHTS CENTRE

Embassy
of the Federal Republic of Germany
Tallinn