Вісник Львівського університету. Серія філос.-політолог. студії. 2023. Випуск 46, с. 373–372
Visnyk of the Lviv University. Series Philos.-Political Studies. Issue 46, p. 373–372

УДК 32.019
DOI https://doi.org/10.30970/PPS.2023.46.47

SOCIAL MEDIA CONTENT MODERATION: THE CASE OF RUSSIA'S WAR AGAINST UKRAINE

Khrystyna Yuskiv
Lviv Polytechnic National University,
Institute of Applied Mathematics and Basic Sciences,
Department of international Information,
Stepana Bandera str., 12, 79013, Lviv, Ukraine

The extent to which russia's war against Ukraine is performed and publicized online is noteworthy. It demonstrated how social media is changing the way war is reported, experienced, and properly understood. Social media posts have evolved into a key information source for both traditional media and open-source intelligence gatherers. Governments may make use of social media as a "tool" to further military objectives. Making sure the world receives information about Ukraine straight from Ukraine is one of the goals.

Platforms explain the blocking of content related to russia's war primarily due to violations of community standards. Algorithms of social media more often mark the content of Ukrainian users as "sensitive" (for example, photos of the bodies of civilians and (or) military personnel). At the same time, the publication of such information helps to document the consequences of war crimes, to raise the awareness of the world community, contribute to the provision of international financial, humanitarian, legal and military aid to Ukraine.

Using keywords, which are prohibited by the community standards regardless of their value, relevance, or significance, is the primary cause of social media post blocking. The practice of combating so-called "border content", where materials are not shown in news feeds but are not blocked, has become common.

It is crucial to bring in unprejudiced evaluators who have knowledge of the context and appropriate cultural and political background. We also consider it permissible to partially revise the requirements of the standards regarding the prohibition of "hate speech". Social media should also take the context of each individual case into account.

Key words: social media, community standards, content, moderation, hate speech, sensitive content, blocking.

Statement of the problem. Being a very powerful tool for the formation of public opinion, with the beginning of russia's military aggression against Ukrane in 2022, social network algorithms began to strongly limit content about the russian-Ukrainian war and block the accounts of public activists who publicized the war crimes of the russian army on the territory of Ukraine. That is, the platform did not become a neutral intermediary in mass communication and began to make biased decisions more often. Social media platforms block pro-Ukrainian content by making reference to community standards that guarantee the protection of network users' safety and psychological health from violence and criminal activity, as well as unacceptable content [1].

Analysis of the latest scientific research. Various aspects of content moderation by social networks were the subject of scientific analysis by domestic and foreign researchers, such as O. Frolova, S. Lyulko, M. Zinchenko, M. Kearney, G. Michael, D. Skrynka. Among domestic scientists it is worth mentioning T. Avdeeva, S. Fiyalka. At the same time, the issue of content blocking by social networks in wartime requires further clarification.

The purpose of the study is analysis of the reasons for the removal of pro-Ukrainian content by algorithms and moderators of social networks, in particular Meta, to identify community standards that need to be revised in wartime conditions.

Presentation of the main material. Social networks use two different modes of moderation to identify content that does not adhere to their truth policy: the first is manual, and the second is automated with artificial intelligence and algorithms, allowing it to process all of the content present on the platforms, be it text, photos, or videos, and analyze content containing external links. If a post does not comply with the platform's security policy, it will be removed or blocked. Front-line personnel that process reports and review content are not the only ones who provide social media moderation services. Journalists, fact-checkers, and academics: the role of these professionals is to support networks in developing their moderation policies, in identifying and qualifying content that is disinformation, and in understanding its characteristics to ensure detection [2].

In total, the Meta Community Standards contain six sections. This is a set of rules that govern what is allowed and what is not allowed on this social network. The standards prohibit: violence and incitement; fraud and deception; intimidation and humiliation; violence and its illustrations; sexual activity; cruelty; and intellectual property. Usually, only after the user submits a complaint does the moderator decide whether to unblock him or leave him blocked [3].

Social media content moderation processes have traditionally been very opaque. Particularly, the above-mentioned Facebook community standards that defined what content was unacceptable were superficial and ambiguous: although they made it clear that "terrorist content" and "hate speech" were prohibited, users had no idea what exactly fell under each of these headings until relatively recently. Facebook's standards are based on feedback from users of the social network and advice from experts in fields such as technology, public safety, human rights, and more. At the same time, content that is potentially subject to prohibition, but of significant public interest is allowed to be published [1].

The standards declare the protection of information consumers against violence and criminal behavior; protection against objectionable content and assurance of authenticity and privacy are guaranteed. Facebook removes wording that incites violence and also "in an effort to prevent and disrupt real-world harm, we do not allow organisations or individuals that proclaim a violent mission or are engaged in violence to have a presence on Facebook" [3].

Meta defines "hate speech" as violent or dehumanizing language, harmful stereotypes, statements of inferiority, expressions of contempt, disgust, or contempt, profanity, and calls for exclusion or segregation [1]. However, the terminology used in hate speech is a subject of some debate. The Meta company recognizes [4], that regional and linguistic peculiarities of word usage are important in the context of a geopolitical situation when the same concepts can be interpreted both neutrally and as hate speech [5].

A new wave of content blocking began with the invasion of the russian federation on February 24, 2022. Ukrainians produced a large amount of content, which is not only a reflection of all events, but also a method of dealing with information. However, even before the invasion, there were precedents for restrictions on some users. For example, on February 23, 2022, Twitter said it mistakenly suspended [6] the accounts of several OSINT reporters who wrote about the buildup of russian military forces near Ukraine.

Meta attempted to modify a content policy based on the idea of political neutrality in light of the ongoing war in Ukraine. At the same time, since April, Instagram has restricted some hashtags for not fitting the Community rules, including #BuchaMassacre, #Azov, #RussiaIsATerroristState, and #StandWithUkraine. As posting violent scenes is prohibited by

the platform's policies, posts using these hashtags were immediately deleted or blocked. In June 2022, the situation repeated itself with the hashtag #russiaisaterrorisstate. Users published content with this hashtag that proved systematic targeted attacks by the russian army on civilian buildings with people inside. To bypass the restrictions, Ukrainians changed the hashtag to #russiaterroristate (without the "is"). In fact, users could publish posts with that hashtag, but it didn't show up in other users' feeds. Meta claimed that they did not impose special sanctions. Algorithms limited the display of this hashtag due to complaints about it [5]. War victims are being blocked, shadowbanned, or silenced.

In our opinion, social networks should regulate freedom of expression in a balanced way during such periods. Many platforms made decisions that were at odds with earlier regulations. The first attempts in this direction were made by Meta. For instance, Facebook altered its hate speech policy in March, 2022 and let users in Ukraine to write postings encouraging violence against "russian invaders" (breaking with the company's general rule that prohibits users from sharing such content), publish videos and photos with the "Azov" regiment, etc. However, later on the network, one could notice quite a lot of hesitation on the part of company representatives and statements that the new approach was misinterpreted, and hate speech should be prohibited in any form [7]. For example, the account of blogger P. Nek was blocked for publishing photos of the dead from Bucha and photographic evidence of systematic, aimed shelling of civilian buildings by the army of the russia. In September 2022, a post with a photo of a fragment of a body exhumed in Izyum was blocked ("Potentially unacceptable content; this photo may contain scenes of violence or content that is difficult to understand"), an illustration by O. Grekov, who depicted a hand with a blue-yellow bracelet ("This photo does not violate community norms, but may contain content that is difficult to perceive") [1].

It is also important to note that as soon as the platforms start to restrict or delete specific content - the Streisand effect is triggered (a phenomenon in which an attempt to remove some information, people start spreading it even more). This is what happened with the quote of Roman Ratushnyi, a Kyiv activist who died at the front. Facebook and Twitter deleted user posts containing a quote: "The more russians we kill now, the fewer russians our children will have to kill" (the system is set to perceive it as hate speech) [8].

The content of illustrators and artists (pictures, gifs, videos, etc.) is much less often removed from social networks, even if such content contains words included in the list of "triggers". Such art is a way to document reality without censorship. Twitter has begun to block the accounts of Ukrainian activists, volunteers, and journalists, particularly if they host fundraising events for the Armed Forces and war-affected citizens. Numerous Ukrainian users claim that after the war began, when they attempted to post fundraisers for the army, their accounts were suspended [9].

All such blockings have a negative impact on the reach of pages: bans and removal of posts lead to "shadow bans", the essence of which is that any user-generated content is not visible to other users. Getting out of the shadow ban can take weeks or even months. The worst-case scenario for the media is that the account is deleted without the possibility of recovery after several warnings from the platform [5].

The government now uses social media as a tool to gather data, communicate with residents directly, and even lobby for international support. Although using social media by leaders and governments is nothing new, the urgency and difficulties brought about by russia's aggression against Ukraine show the platforms' value in the context of war for gathering and disseminating information about russia's war crimes to large audiences in a timely manner. For instance, the Ministry of Digital Transformation of Ukraine created a Telegram chatbot that enables users to share videos and locations of russian soldiers [10].

The circulation of false information about russia's invasion of Ukraine is a sign of deeper problems related to the transformation of the information landscape. By enabling the creation of echo chambers and confirmation bias processes that separate the news and information users perceive and deal with online, platform and algorithm designs can accelerate the spread of misinformation. People have a tendency to disseminate false information "farther, faster, deeper, and more broadly than the truth", [10] which is especially true of false political news. So, the first level of sanctions on the part of the platforms is that they try to respond in a targeted way by blocking access or removing publications that violate the terms of use. If such measures are not possible, social networks begin to penalize accounts or pages.

For instance, in reaction to the war in Ukraine, social media have started [11] removing pro-russian accounts that incite hate in large numbers. However, such measures are implemented in both directions, and this presents a challenge because Ukrainians' accounts are frequently restricted as well. For example, Meta prohibits [12] the use of the words "rusnya" or "moskal" because it is qualified as hate speech. Given that these characteristics are ubiquitous at the onset of full-scale war, the scale of the blockades is astounding.

It is important to note here that if the content is incorrectly qualified, the decision to block the account can be fatal for the media or influencers. Although blocking appeal mechanisms exist, they are not always effective. In Ukraine, there was a period of blocking #Bucha and #Irpin, when thousands of posts with evidence of war crimes were removed en masse from the platforms. The restriction was imposed due to an alleged violation of the ban on depiction of violence. However, in practice, this led to the blocking of hundreds of accounts [7]. This case shows the need for a more careful evaluation of restrictive measures in the information field, which is filled with information about armed conflict.

Borderline content removal, which is not subject to be banned but might "deteriorate the quality of services", was one of Facebook's policies. As a result, Facebook has taught its AI algorithms to recognize this material and share it less (shadow banning) [1].

Even though the russian-Ukrainian war is still going on, social media administrators are still unsure of how to respond to emotional posts and horrific pictures posted by Ukrainian users. Occasionally, the problem could be solved only through a personal appeal of the Minister of Digital Transformation of Ukraine to the management of the social network; this was the case, in particular, with the Instagram account of the Association of Families of Defenders of Azovstal [13].

It is worth summarizing that, as of February 24, 2022, social platforms have two options for action. The first and easiest is to follow the old rules about hate speech and "borderline" content. These rules, for example, provided for the blocking of all mentions of the Azov regiment, the banning of any calls to violence, and other moments that became commonplace in social networks after February 24.

Another problem was the spread of propaganda by the russian state media. In fact, social networks have found themselves in the crossfire of content that, on the one hand, contradicts itself and, on the other hand, partially violates the rules of the service. Companies could resort to the universal policy of banning all war content. On the one hand, it made moderating easier, but it also needed a lot of work to the moderation algorithms. Another option is to change the rules for a long time, or at least for the duration of the military confrontation. Or make a decision to block the aggressor's propaganda posts while relaxing the posting rules for the defending party. In fact, most social services have chosen this path.

The Meta has become one of the most radical and quickest in decision-making regarding this war. A few days after the start of russia's invasion, Meta removed a large number of fake Facebook and Instagram pages and groups that attacked Ukraine. In the history of social networks,

there has never been a decision to change the policy regarding hate speech and death wishes for the russian military. The automated censoring of pictures from Bucha in early April 2022, which was carried out using image recognition algorithms, served as at least one example of the need for such adjustments. However, Ukrainian journalists complain that Meta moderators often do not follow this rule. This caused a call from Ukrainian public organizations to make appropriate changes to the Meta rules for military content.

However, social media have not completely solved the problem of content moderation in difficult conditions. This was shown, for example, by the complaints of blocked users due to old posts, and the automatic reaction to photos from Bucha/Irpin or any another city of Ukraine. As well as blocking posts with the hashtags #standwithukraine or #russiaisaterroriststate. Moreover, a large number of popular users who actively used this hashtag got a "shadow ban". Facebook could impose such sanctions on users and hashtags automatically due to the flow of complaints. And to cancel it, you need to study the situation and have someone from the team of moderators make a decision, which obviously takes time [14].

Back in July 2022, the Ministry of Digital Transformation of Ukraine appealed to Meta to improve the quality of moderation, not block Ukrainian bloggers, and not hide content. According to the Ministry, part of the content on Facebook and Instagram is blocked by the automatic algorithms of social networks [15]. In response, Meta recently developed and adopted an adapted content moderation policy for war coverage. Gradually, this decision begins to take effect. In particular, Meta will exclude the Azov regiment from the blocking policy. Content posted by others regarding the Azov Regiment will no longer be deleted (if it doesn't violate Meta policy) [16]. This is an important step and a signal to the global community that the truth lies with Ukraine.

It appears that platforms have not properly stated the basis on which they have reached their judgments. Or how it would apply in other situations where there is a lack of a clear decision-making structure or more transparent efforts to gather external experts to provide context and guidance. Such a strategy runs the risk of allegations of hypocrisy and the impression of inconsistent decision-making. It's important to note that social media corporations were unable to adequately combat the risk posed by russian misinformation prior to the war [10].

Conclusions and prospects for further research. Social networks have their own algorithms that automatically block "unacceptable" content without the involvement of a moderator. For example, it can be unedited photos of civilian or military bodies. The decision of the governments of Western countries to provide aid to Ukraine depends on the mood of the populations of these countries; therefore, for example, the distribution of content that reveals war crimes of russia can contribute to the provision of such aid and vice versa. Algorithms are also the result of the work of people, who can put their prejudices and sympathies into them.

By community standards, all online users are equally protected, but historical context and current events that may lift the taboo on covering sensitive content must also be taken into account. This requires reviewers with relevant cultural and political experience, an understanding of the context, and an awareness that the social media community is, on the one hand, one global community and, on the other hand, a set of different communities with different experiences of interaction with other communities. Social media posts are also a significant source of publicly available data that is being analyzed and even utilized as proof of war crimes. Examples include satellite photographs, videos, and pictures as well as photos or videos of eyewitnesses or victims of russian war crimes.

Regulating the information space during a period of armed aggression is definitely not an easy task. However, war should not become a reason for uncontrolled restrictions on

social networks because sometimes they remain the only resources from which readers can get objective and necessary information. Likewise, one should not expect ideal behavior from the platforms because it is hardly possible to predict absolutely everything and react to each of the threats in time.

References

1. Фіялка, С. В. (2022). Проблематика дотримання стандартів спільноти в соцмережі «Фейсбук» в умовах збройної агресії Російської федерації. Обрії друкарства, 2(12), 5–17. https://doi.org/10.20535/2522-1078.2022.2(12).270922 (дата звернення: 11.02.2023)
2. How do social networks manage content moderation to combat misinformation? URL: https://www.buster.ai/blog/how-do-social-networks-manage-content-moderation-to-combat-misinformation (дата звернення: 11.02.2023)
3. Facebook Community Standards URL: https://transparency.fb.com/uk-ua/policies (дата звернення: 13.02.2023)
4. Hard Questions: Who Should Decide What Is Hate Speech in an Online Global Community? URL: https://about.fb.com/news/2017/06/hard-questions-hate-speech/ (дата звернення: 14.02.2023)
5. Потенційно неприйнятний: як онлайн платформи обмежують контент українці URL: https://ukrainer.net/pravo-na-kontent/ (дата звернення: 14.02.2023)
6. We're closely investigating – but mass reporting is not a factor here. URL: http://surl.li/euzhj (дата звернення: 12.02.2023)
7. Незоряні війни: соцмережі під час збройних конфліктів URL: https://cedem.org.ua/analytics/sotsmerezhi-zbroyni-konflikty/ (дата звернення: 11.02.2023)
8. Facebook і Twitter блокують цитату загиблого на фронті Романа Ратушного про війну з Росією URL: https://ms.detector.media/sotsmerezhi/post/29670/2022-06-15-facebook-i-twitter-blokuyut-tsytatu-zagyblogo-na-fronti-romana-ratushnogo-pro-viynu-z-rosiieyu/ (дата звернення: 13.02.2023)
9. Twitter Bans of Ukrainian Activists Face Pushback URL: https://www.thestreet.com/investing/is-twitter-banning-ukrainian-activists (дата звернення: 11.02.2023)
10. OECD, Disinformation and Russia's war of aggression against Ukraine: Threats and governance responses, OECD Policy Responses on the Impacts of the War in Ukraine, URL: https://www.oecd.org/ukraine-hub/policy-responses/disinformation-and-russia-s-war-of-aggression-against-ukraine-37186bde/. (дата звернення: 13.02.2023)
11. Guns, tanks and Twitter: how Russia and Ukraine are using social media as the war drags on. URL: https://theconversation.com/guns-tanks-and-twitter-how-russia-and-ukraine-are-using-social-media-as-the-war-drags-on-180131 (дата звернення: 11.02.2023)
12. Мова ворожнечі в соцмережах: як регулювати? URL: https://cedem.org.ua/analytics/mova-vorozhnechi-v-sotsmerezhah/ (дата звернення: 11.02.2023)
13. Як російська війна в Україні впливає на модерацію контенту в соцмережах URL: https://glavcom.ua/digest/jak-rosijska-vijna-v-ukrajini-vplivaje-na-moderatsiju-kontentu-v-sotsmerezhakh-891123.html (дата звернення: 11.02.2023)
14. Як російська агресія змінила пошукові сервіси та соціальні мережі URL: https://chas.news/current/viina-v-ukraini-pokazala-scho-neitralitetu-zaraz-ne-mozhe-isnuvati-tehnogigantam-takozh-dovelosya-vibrati-odnu-zi-storin-konfliktu (дата звернення: 11.02.2023)
15. Мінцифри звернулося до Meta щодо приховування українських постів та блокування акаунтів URL: https://espreso.tv/mintsifri-zvernulosya-do-meta-shchodo-prikhovuvannya-ukrainskikh-postiv-ta-blokuvannya-akauntiv (дата звернення: 11.02.2023)

16. Meta пообіцяла не блокувати дописи про полк "Азов" у своїх соцмережах URL: https://espreso.tv/meta-poobitsyala-ne-blokuvati-dopisi-pro-polk-azov-u-svoikh-sotsmerezhakh (дата звернення: 11.02.2023)

ПОЛІТИКА МОДЕРУВАННЯ СОЦІАЛЬНИМИ МЕРЕЖАМИ КОНТЕНТУ: КЕЙС ВІЙНИ РОСІЇ ПРОТИ УКРАЇНІ

Христина Юськів
Національний університет «Львівська політехніка»,
Інститут прикладної математики та фундаментальних наук,
кафедра міжнародної інформації
вул. Степана Бандери, 12, 79013, м. Львів, Україна

Агресивна війна Росії проти України примітна тим, наскільки вона ведеться та поширюється в Інтернет просторі. Повномасштабне вторгнення Росії в Україну проілюструвало як соціальні медіа змінюють спосіб хронікування, переживання та розуміння війни. Дописи в соціальних мережах стали важливим джерелом інформації як для збирачів відкритих джерел інформації (OSINT), так і для звичайних ЗМІ. Соціальні медіа можуть використовуватися як «інструмент» для досягнення урядами цілей військового часу. Частково мета полягає в тому, щоб світ отримував інформацію про Україну безпосередньо з України.

Повідомлення користувачам про блокування контенту, пов'язаного з війною Росії в Україні, передусім стосуються порушення стандартів спільноти. Алгоритми все частіше позначають контент від українських користувачів як «чутливий» (наприклад, фотографії загиблих цивільних і/або військових), а оприлюднення таких матеріалів допомагає повідомити про наслідки військових злочинів і підвищити міжнародну обізнаність. Вони сприяють забезпеченню міжнародної фінансової, гуманітарної, правової та військової допомоги Україні.

В основному, публікації у соціальних мережах блокують через використання заборонених спільнотою ключових слів. незалежно від того, яку користь, суспільний інтерес та сенс він має. Поширеною є практика боротьби з так званим межовим контентом, коли матеріал не показується в стрічках новин і не блокується.

Доцільно залучати неупереджених рецензентів, які мають відповідний культурний та політичний досвід, розуміють контекст. Ми також вважаємо, що можна частково змінити вимоги стандартів щодо заборони «мови ворожнечі». Соціальні мережі також мають враховувати контекст окремих випадків.

Ключові слова: соціальні мережі, стандарти спільноти, контент, модератор, мова ворожнечі, чутливий вміст, блокування.