ISBN 978-9934-619-35-9

Impact of the Digital Services Act

A Facebook Case Study

PREPARED AND PUBLISHED BY THE

NATO STRATEGIC COMMUNICATIONS CENTRE OF EXCELLENCE

ISBN: 978-9934-619-35-9

Authors: Tetiana Haiduchyk, Marat Khusnutdinov, Fabio Belafatti, Jakub Kubś, Andrius Šuminas,
Dr. Gundars Bergmanis-Korāts
Project Manager: Mārtiņš Pundors
Contributors: Trementum Research, Debunk.org
Content Editor: Merle Anne Read
Design: Inga Ropša

Riga, May 2025

NATO STRATCOM COE
11b Kalnciema iela,
Riga, LV1048, Latvia
stratcomcoe.org
@stratcomcoe

This publication does not represent the opinions or policies of NATO or NATO StratCom COE.
The views expressed here do not represent the views of NATO.

Impact of the Digital Services Act

A Facebook Case Study

Contents

Executive summary

The adoption of the EU Digital Services Act[1] (hereinafter the DSA) aimed to create a safe online information environment, specifically within the EU/EEA area. The aim of this research was to measure the effects of the DSA in curbing the spread of harmful content on social media. As measuring the results of such a broad goal was challenging, our study focused on one of the dominant social media platforms: Facebook. To assess the impact of the DSA, we compared the share of harmful content[2] published by Polish and Lithuanian accounts on Facebook before and after the DSA entered into force.

Our multi-stage approach involved using a small AI model, GPT-4o mini, to initially flag harmful content, followed by applying larger models for validating and in-depth reasoning. In total we classified 959 harmful posts from 2023 and 1,392 posts from 2024.

Hate speech targeting individuals based on protected characteristics remains the platform's most significant vulnerability in combating harmful content. About 90% of such content for both languages was related to hate speech in both 2023 and 2024. However, differences in the numbers of posts with harmful content were noted between the two languages. In 2024 such posts in Polish increased by 55%, while those in Lithuanian declined by 11%.

Differences between the languages were also observed in the topics of harmful content, which may be influenced by varying model performance across the languages. However, it should be noted that the study transparently shows the accuracy metric in the *Appendix*. While antisemitism mostly related to the Israel–Hamas war was the most prevalent in both languages, the number of harmful messages related to this topic increased only in Polish in 2024. Therefore, we could argue that it is crucial for platforms to dedicate more efforts to monitoring and responding to the content associated with ongoing political conflicts, since those are essentially a fertile ground for further hostilities online.

At the same time, the increase in harmful posts varied significantly by account type in Polish. Specifically, individual Polish Facebook accounts published 6% more harmful posts, while Polish groups saw an increase of 128% in these posts in 2024. Such a noticeable difference likely indicates the platform's system vulnerabilities in detecting and addressing harmful content within groups.

An assessment of the platform's efforts to ensure a safe information environment before and after the implementation of the DSA showed dual results. On the one hand, we noted that the share of fact-checked posts increased significantly in 2024, likely suggesting an additional platform investment in this area. On the other hand, we tracked the decline in the removal rate of harmful posts, which was particularly sharp for Lithuanian content. This may indicate insufficient efforts to monitor and remove violations involving small languages.

As a result, the study demonstrated that despite certain improvements the platform made in creating a safe online environment, we could not claim an overall enhancement after DSA enforcement. Additionally, our results highlight the current vulnerabilities and areas for improvement that the platform should address.

This research is just a glimpse into a field of research with an expansive scope. Therefore, our conclusions and recommendations should be seen as an encouragement for further and wider research across online social media platforms.

Introduction

Social networks are becoming an integral part of society today, which requires them to continuously combat increasing risks and threats online. Global challenges such as the COVID-19 pandemic, the Russian invasion of Ukraine, and attempts to meddle in political elections around the globe have demonstrated vulnerabilities of the digital environment to manipulation and the spread of disinformation. In response the European Union adopted the Digital Services Act in 2022, which introduced a regulatory framework for online platforms to monitor and safeguard the online information environment. As of 17 February 2024, the DSA rules applied to all online intermediary services providers that offer their services in the EU/EEA, including the most popular social media platforms.

This study was designed to measure the effectiveness of DSA enforcement on Facebook due to its dominance in the EU. We compared the spread of harmful content among Lithuanian and Polish Facebook accounts before and after the DSA entered into force. The multi-stage approach utilising advanced AI models allowed us to identify and classify harmful content. We conducted a comparative analysis focusing on trends in the spread of harmful content and the impact of the DSA on these trends. Furthermore, we examined the platform's efforts to limit the dissemination of such content based on an assessment of independent fact-checker involvement and the removal rate of harmful posts.

Methodology

Data

Our case study of harmful content prevalence online before and after the DSA focused on Facebook, and we looked for any changes in content since the DSA came into force. We compared two datasets of posts from 2023 and 2024, namely data from Lithuanian and Polish Facebook accounts.

In this research we explored the content of Facebook accounts that were most likely involved in posting harmful content. Such Facebook groups and pages were identified with the assistance of investigative journalists and fact-checking organisations from Lithuania and Poland. In Lithuania, the list of disinformation pages and groups was compiled by journalists from the national broadcaster Lithuanian National Television and Radio (LRT) in collaboration with Debunk.org. This investigation focused on a pro-Kremlin network of connections on Facebook.[3] In the case of Poland, the list of Facebook pages and groups linked to disinformation was derived from Debunk.org's collaboration with journalists from the non-governmental organisation VSquare, as part of an international project analysing disinformation networks on social media and blogs in Central and Eastern Europe.[4]

Table 1 shows the initial data collected during the two stages of the research.

	Timeframe	Number of posts	Number of accounts
2023	25 May – 25 August	405,232	1,429
2024	1 March – 31 May	230,031	1,282

TABLE 1. Overview of the initial data

In the 2024 dataset we observed that some accounts were either unavailable, set to private, or inactive, compared to the 2023 dataset. To ensure comparability of both datasets, we focused exclusively on active accounts from both datasets, excluding inactive, unavailable, and private accounts from the previous dataset. Consequently, the comparative analysis was based on 1,180 Polish and 102 Lithuanian accounts. It is important to note that the difference in sample size reflects the relative size of the language groups, with Polish representing a larger language population compared to Lithuanian. Figure 1 displays the status of accounts in 2024 by country.

posts from an account discussing the war in Ukraine, as a specific topic might be more likely to contain harmful content. To ensure comparability, we normalised the dataset by matching the number of posts for each account in both datasets. For accounts with different post volumes in 2023 and 2024, we randomly reduced the larger group to match the smaller one, creating a balanced dataset for accurate year-on-year analysis.

This approach ensured the comparison of equivalent data volumes, consisting of **165,201 comparable pairs** (330,412 posts in total) **from 1,274 accounts**, including 22,029 posts from 100 Lithuanian accounts and 143,172 posts from 1,174 Polish accounts. In this study we considered the textual components of posts, including the text of original and reposted posts, link preview text, and – by utilisation of optical character recognition (OCR) – text from images.

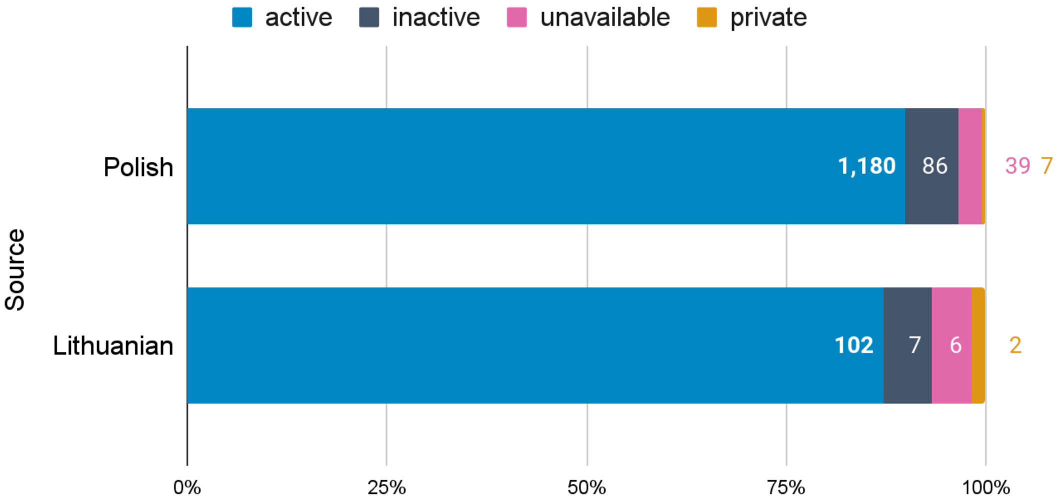

FIGURE 1. Number of accounts by status

The 2024 dataset consists of accounts publishing on various topics, such as Polish and Lithuanian right-wing groups and individuals promoting nationalism, social conservatism, anti-establishment views, historical revisionism, and conspiracy theories. Thus, it is not appropriate to equate 100 posts from an account focused on religion with 100

Detecting harmful content

Facebook claims[5] that the platform's technology detects and removes the vast majority of violating content even before users report it. Meanwhile, potentially violating content detected by the technology is sent to review teams for evaluation and action. We applied a multi-stage approach (see Figure 2) to detect potential violations in a limited

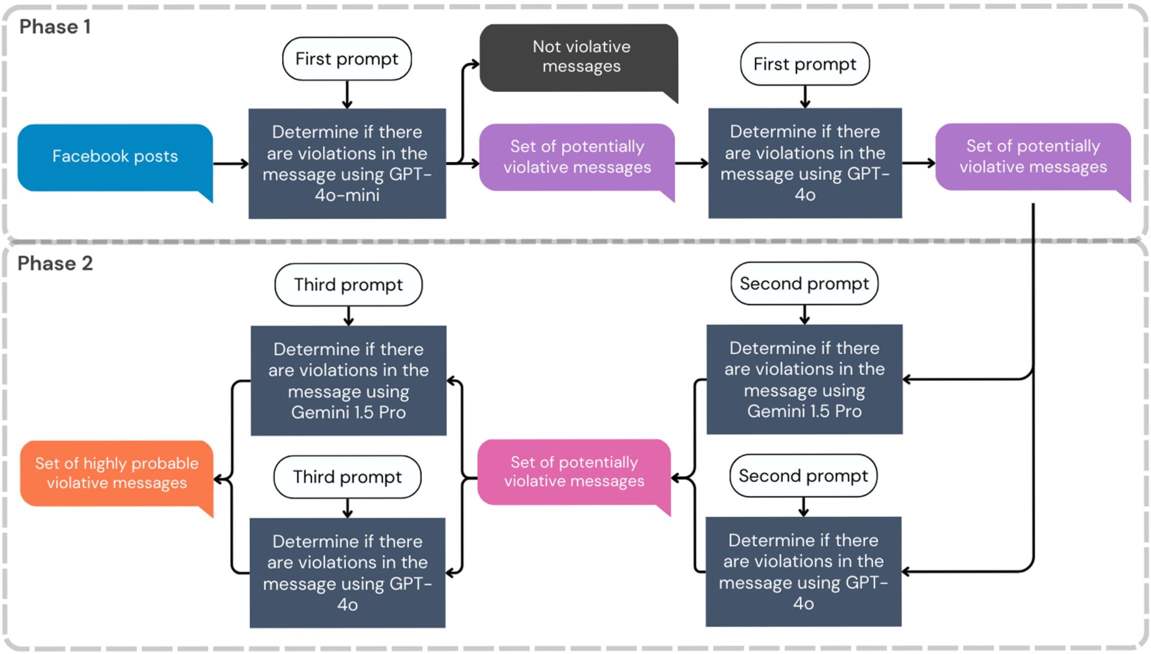
Phase 1
Facebook posts
First prompt
Determine if there are violations in the message using GPT-4o-mini
Not violative messages
Set of potentially violative messages
First prompt
Determine if there are violations in the message using GPT-4o
Set of potentially violative messages
Phase 2
Second prompt
Determine if there are violations in the message using Gemini 1.5 Pro
Second prompt
Determine if there are violations in the message using GPT-4o
Set of potentially violative messages
Third prompt
Determine if there are violations in the message using Gemini 1.5 Pro
Third prompt
Determine if there are violations in the message using GPT-4o
Set of highly probable violative messages

FIGURE 2. The multi-stage approach to detect potential violations

dataset of Facebook posts described above as a first stage.

For Phase 1 we created a unified set of 14 rules based on the DSA and Facebook Community Standards policy.[6] Using these rules, we built a complex *prompt* for AI models to detect potential violations in the provided text. In this study we will use the term ‘harmful content’ to refer to content potentially containing violations. Utilising this prompt, the following steps were undertaken:

1. Initial filtering for all posts using the GPT-4o mini model was applied to flag whether a specific rule from the list had potentially been broken.

2. The subset of potentially harmful posts classified by the GPT-4o mini model was passed to the most capable GPT-4o model at that time with the same prompt as in Step 1. This step reduced the number of falsely flagged harmful posts and assigned more correct labels.

Phase 2 included cross-validation of violations by two AI models. We chose this approach to eliminate the possibility of a single-model bias and minimise the number of false positive results. In addition to the GPT-4o model, we used the Gemini 1.5 Pro as a comparable flagship model. According to benchmark results,[7] these two models are on a par in terms of the Artificial Analysis Quality Index.[8] This index is a metric used to evaluate the performance and reliability of AI models, ensuring their effectiveness and trustworthiness in analysis results.

We applied GPT-4o and Gemini 1.5 Pro models at this stage. If both models independently identified the same rule violation within a post, we considered the rule to be broken. If only one model flagged a potential violation while the other did not, the rule was considered not violated.

In Phase 3 we identified the rules where AI models were most likely to make mistakes. We then further processed these rules. The rules with few true positive results and high false positive rates were removed at this stage. This choice was made to improve the overall accuracy of the AI models and reduce the risk of incorrect classifications. We combined the “discrimination rule” with the "hate speech rule" and processed only this combined rule with the GPT-4o and Gemini 1.5 Pro models in the same cross-validation manner as in Phase 2.

To assess the reliability of the approach used, we manually marked 50 posts (or the maximum possible number of posts if fewer than 50 for a specific rule) for each of the 13 violated rules and for each country (1,295 posts in total) that the GPT-4o mini model labelled as violative according to a specific rule. Moreover, not only was the presence of a violation annotated, but compliance with a specific rule was also noted. If, during manual labelling, a post contained a violation that corresponded to a rule different from the one specified by the AI, it was labelled as non-compliance with that specific rule.

These manually labelled posts allowed us to evaluate the accuracy of the AI labelling in Phases 1, 2, and 3 and identify where AI models made mistakes in labelling. This evaluation helped us adjust the prompt to improve results based on the accuracy of the responses related to specific rules.

Facebook efforts

Independent fact-checking is one of the primary tools for mitigating risks associated with disseminating disinformation. Thus, assessing the platform's efficiency in labelling disinformation is essential when evaluating the implementation of the DSA. In this study, we compared the proportion of posts labelled as fact-checked and published by analysed Facebook accounts before the DSA implementation (based on data from 2023) and after its implementation (based on posts published in 2024).

Another metric we used in this study to measure the efficiency of the DSA was the proportion of harmful content that the platform removed or disabled access to. A lack of efforts to remove or disable such content may indicate insufficient enforcement practices on the part of the platform, raising questions about the overall effectiveness of the DSA.

Harmful content comparison

According to our analysis, harmful posts increased by 45% within analysed accounts after the DSA implementation. Over 90% of such posts in both years were related to hate speech posts, which had a major impact on the final results. However, we noticed different patterns between the Polish and Lithuanian languages: in 2024 there were 55% more potentially harmful posts in Polish, whereas 11% fewer such posts in Lithuanian. Even though antisemitism was the most common issue in both languages, only in Polish did the number of violent messages on this topic increase in 2024.

The detection of harmful content involved two main phases: initial screening by sequentially applying GPT-4o mini and GPT-4o to comparable datasets, and cross-validation with parallel use of GPT-4o and Gemini 1.5 Pro. By simultaneously applying larger models, we increased the accuracy from 58% to 70% and reduced the false positive rate from 38% to 7%. Please refer to the *Appendix* for more details on the methodologies and results. In total 2,351 Facebook posts were classified as harmful by AI models across both languages.

Violation types

We detected a significant increase in harmful content published in 2024 (1,392 posts) compared to 2023 (959 posts). Notably, over 90% of the posts classified as harmful by AI were related to hate speech in both years. At the same time we observed a significant increase in posts with this type of violation – 49% more cases consisting of hate speech in 2024 (Figure 3).

Categories such as public safety, illegal products, and election integrity also prevailed in 2024. In contrast, categories like threats against officials and terrorist content decreased in 2024. Please refer to the *Appendix* for examples of these posts for each rule.

Therefore, it is essential to highlight that, according to our study, this rise in harmful content after the DSA implementation is primarily attributed to hate speech, which not only represents the largest category but also shows the most significant growth. However, if we exclude hate speech posts from the comparison, the number of harmful posts in 2023 and 2024 is about the same.

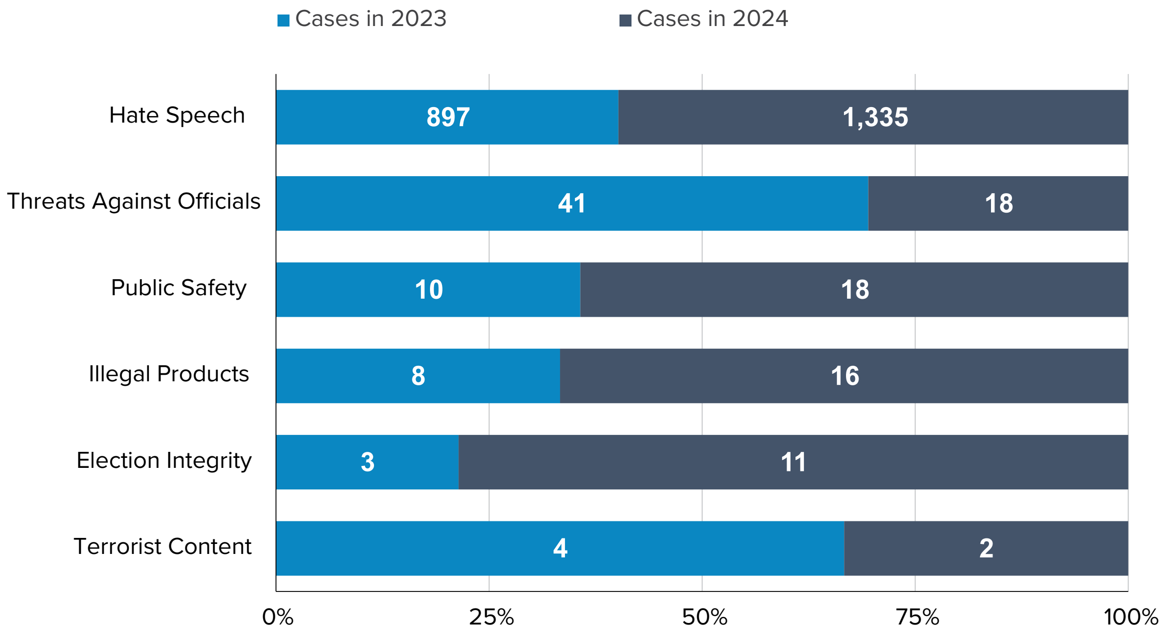

FIGURE 3. Posts classified as harmful, by type of violation

Language differences

Although we observed an overall 45% increase in harmful posts in 2024, the results varied between Lithuanian and Polish posts. As Figure 4 shows, the number of harmful posts in Polish increased by 448 (+55%) in 2024, while in Lithuanian it decreased by 15 (−11%).

Furthermore, we observed differences across the languages not only in the overall change in harmful posts but also in specific types of violations (Figure 5). Among Polish accounts the number of posts containing hate speech increased by 456, representing a 59% rise compared to 2023. The situation in Lithuanian sources was different, with the posts marked as hate speech decreasing by 14%.

The decrease in harmful posts in the Lithuanian language may be a result of improved Facebook content moderation after the DSA enforcement. However, more detailed monitoring is necessary as our study is limited to only two data periods, and thus we have provided a yearly comparison within the scope of this research.

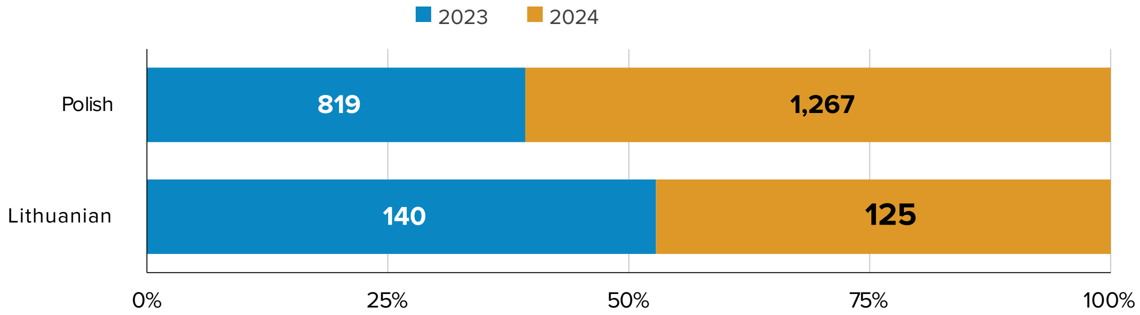

FIGURE 4. Posts classified as harmful, by language

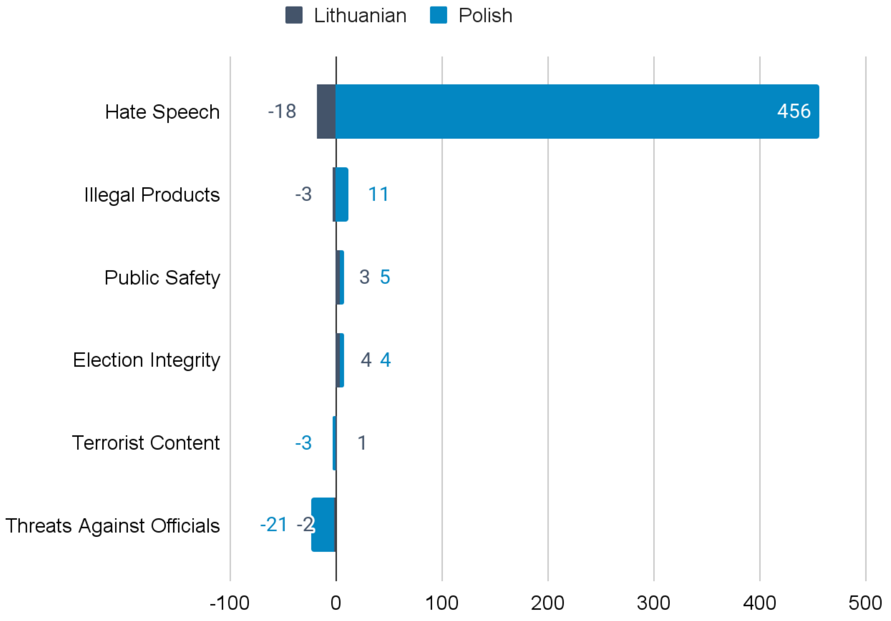

FIGURE 5. Change in harmful content across languages, 2023 vs 2024

Concurrently, the increase in harmful posts in the Polish language varied significantly according to account type (Figure 6). While Facebook accounts posting in the Polish language published 6% more of such posts, we detected 128% more harmful posts published in Polish language groups in 2024. This highlights a critical issue that needs to be examined in future research: the potential differences in flagging mechanisms regarding their approaches to detecting and addressing content violations across various account types.

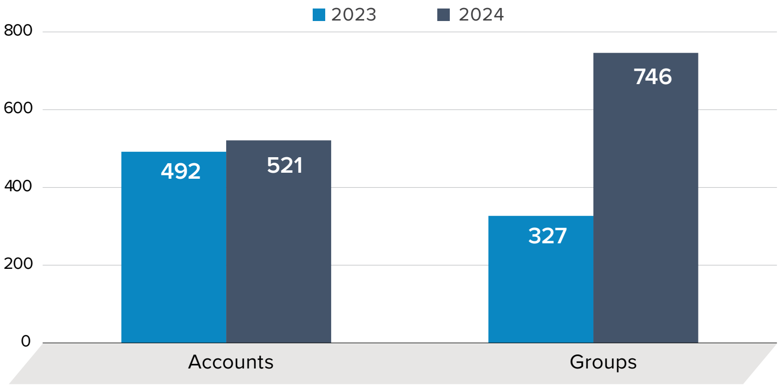

FIGURE 6. Polish posts classified as harmful, by account type

Shifts in harmful posts by topic

This section demonstrates the changes in harmful posts by topic between 2023 and 2024. As Figure 7 shows, the number of harmful posts increased significantly in 2024 among the most common topics, such as antisemitism and anti-EU and anti-globalist sentiment. However, this increase was observed only for Polish content. Among Lithuanian posts, we observed fewer antisemitism posts in 2024, while growth was observed in posts on anti-LGBT sentiment.

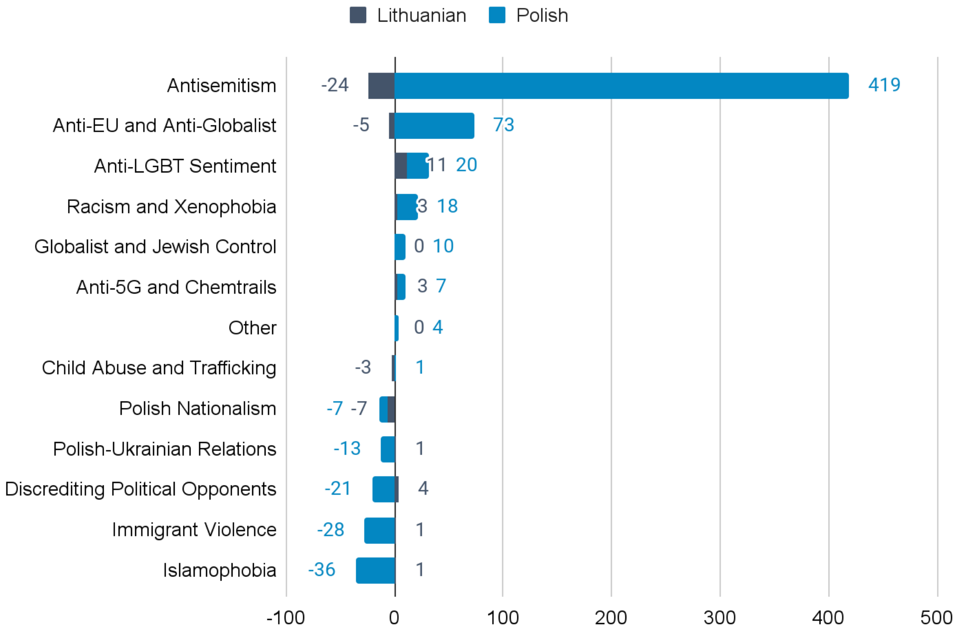

FIGURE 7. Change in the number of harmful posts, by topic

It should be noted that the topic of antisemitism was the most common for both languages, accounting for 41% of posts classified as harmful. At the same time, most of the antisemitism posts were related to the Israel–Hamas war that began in October 2023, suggesting possibly insufficient measures by the platform to respond to digital threats related to the conflict.

Facebook efforts

We observed different effects of the platform's steps in such areas as the involvement of independent fact-checkers and the removal of harmful content. The proportion of fact-checked posts increased for both languages after the DSA implementation, suggesting additional platform investment in creating a safe online environment. At the same time, decreasing the removal rate of harmful posts likely indicates insufficient efforts, which has a particularly negative impact on small languages.

Independent fact-checker involvement

Facebook claimed that the platform partners with independent third-party fact-checkers to counter the spread of disinformation.[9] The DSA aims to create a safer online environment where countering disinformation is a critical step. The understanding of independent fact-checker involvement is essential in evaluating the effectiveness of the DSA, as it shows the actual investment of the platform in creating a safe online environment.

We compared the number of posts labelled as fact-checked within analysed Facebook accounts. In 2023 we found 347 fact-checks for 165,201 posts, compared to 502 for the same number of posts in 2024. The change between the two years is 155 additional fact-checks or 45% more fact-checked posts in 2024. Additionally an increase in the number of fact-checked posts was observed for both Polish and Lithuanian (Figure 8).

We observed increased fact-checked posts in both languages, unlike harmful content. That is why, even considering the growth of harmful content demonstrated in the

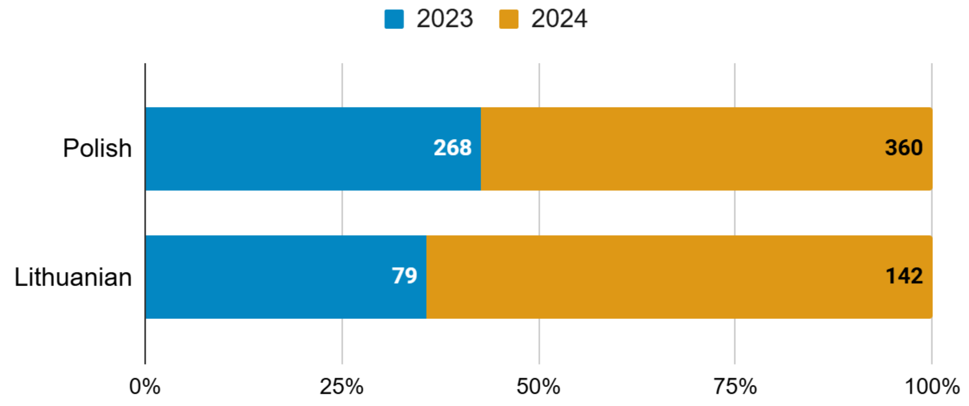

FIGURE 8. Number of fact-checked posts per language

previous sections, the increase in the number of fact-checked posts is less likely to indicate an actual rise in false information on the platform. Moreover, this trend likely reflects the platform's enhanced investment in fact-checking efforts during the research period. However, in early 2025 Meta announced the end of its third-party fact-checking programme and a move to a 'Community Notes model'.[10] This trend shows the platform's reduced willingness to invest in combating disinformation, which may raise concerns about the effectiveness of disinformation detection and the overall safety of the platform.

Removal efforts

Platforms cannot control users' intentions to spread dangerous content, but they are obliged to respond to already published content that would violate any of the existing rules and policies. In particular, the DSA obligates platforms to remove or disable access to such content. This section will assess the platform's efforts in this direction by comparing potentially harmful content classified during this study.

The availability of harmful posts was checked in November 2024 for both datasets. That means that the posts from the 2024 dataset were reviewed at least 5 months after publication, and those from the 2023 dataset, 1 year and 5 months after the publication. We consider such a period enough for the platform to take the appropriate measures to counter harmful content.

We observed an overall decline in the removal of harmful posts between 2023 and 2024. At the time of our analysis, 12% of harmful posts published in 2023 were unavailable. In comparison, the proportion of removed posts was significantly lower for those published in 2024, with only 4% of harmful posts being unavailable.

Although both languages saw a decrease in deleted harmful posts, the dynamics across the languages were slightly different (Figure 9). The share of removed Lithuanian harmful posts from 2023 decreased significantly from 33.3% to 0.8% from 2024. For Polish content, this percentage changed from 11% to 4.5%.

Considering that only one post published in 2024 was removed among Lithuanian harmful posts, this indicates insufficient efforts by the platform, which could be an especially acute issue for small languages.

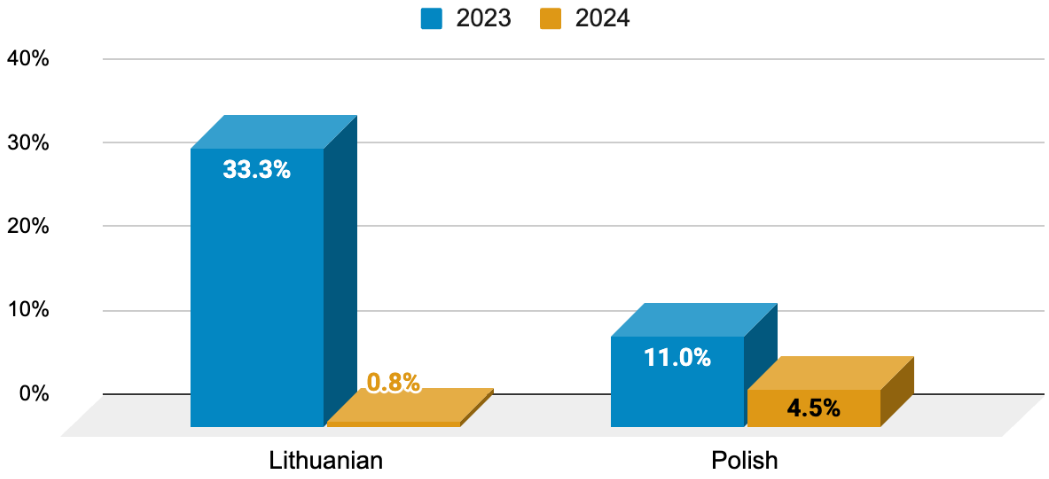

FIGURE 9. Share of unavailable posts classified as harmful

Conclusions and recommendations

While we acknowledge the methodological constraints of this study – including a limited sample size and variability in sample parameters (such as fluctuations in user account availability and user group activity) – the findings provide valuable preliminary insights. These indicative outcomes underline the importance of additional research to substantiate and further explore the observed phenomena across broader and more diverse contexts on various social media platforms.

Insufficient legislative impact

Although measuring the effectiveness of the DSA enforcement requires a comprehensive approach that includes evaluations across multiple platforms, we believe that assessing one of the dominant very large online platforms (VLOPs) is an important indicator of the legislative impact. The Facebook case study shows a 55% increase in harmful posts published by Polish accounts in 2023 and 2024, suggesting a possible lack of efforts to safeguard the online environment after the introduction of the DSA norms.

Hate speech targeting individuals based on protected characteristics has proven to be the most vulnerable aspect of the platform's efforts to combat violations. Even among Lithuanian-language Facebook posts, where there has been an overall decrease in harmful content, hate speech continues to account for around 90% of all violations.

Responsibility for violations

We observed the difference in disseminating harmful content not only across the languages but also among different account types. The increase in violations was more significant in Polish Facebook groups than in individual accounts. This finding suggests that the platform is more likely to focus on regulating content associated with clearly identifiable responsible actors. At the same time, it may struggle to regulate content where several responsible parties can be identified – for example, in Facebook groups, the group owners and post authors. This trend threatens the overall safeguarding of the information environment, as Facebook groups are an important part of the platform's ecosystem and encourage a significant number of users.

Limited efforts regarding armed conflicts

Among posts classified as potentially harmful, the topic of antisemitism was the most common for both languages, accounting for 41%. Most of these posts related to the Israel–Hamas war that began in October 2023. The significant proportion of violations related to antisemitism underscores how ongoing geopolitical conflicts can provide an increasingly fertile setting for further hostilities online. As a result, there is a critical need for platforms to dedicate more efforts to monitoring and responding to online threats associated with ongoing armed conflicts.

Fact-checkers' involvement

The comparison between the proportion of fact-checked posts in 2023 and 2024, in our assessment, is a valuable point in the DSA impact assessment, as it indicates the actual platform investment in combating false information during the study period. Positive trends were observed in both languages, with an overall increase of 45% in fact-checked posts. The concurrent growth in the share of such publications in both languages likely suggests that the platform has made additional efforts in this area. However, Meta's recent announcement regarding the end of the third-party fact-checkers programme and shift towards a user-generated content moderation model raises concerns about the effectiveness of disinformation detection on the platform.

Additional challenges for small languages

While the removal rate of harmful posts decreased for both languages, the decline was much steeper for the Lithuanian language. That is, 33% of harmful Lithuanian language posts published in 2023 were removed, while over 99% of the posts published in 2024 remain online five months after publishing. This underscores the urgent need for platforms to invest more in internal monitoring systems and addressing harmful content, particularly in technically less supported languages. Such investments would allow the achievement of balanced content moderation.

Appendix

Harmful content detection

Phase 1: initial screening

According to our methodology, the first phase of this research involved sequentially processing our datasets using two models, GPT-4o mini and GPT-4o, to identify the set of harmful messages. We applied the same prompt for both models; however, GPT-4o processed only those posts that GPT-4o mini identified as potentially harmful.

After applying the GPT-4o mini model, we identified 23,675 potentially harmful posts created in 2023 and 22,614 in 2024. As shown in Figure 10, among Lithuanian posts the GPT-4o mini model labelled 3,857 posts as harmful in 2023 and 3,170 posts in 2024, indicating 18% fewer harmful publications in 2024. Meanwhile, for Polish messages, the share in the two years is nearly identical; there were 19,818 harmful posts in 2023 compared to 19,444 in 2024, representing a decrease of 1.9%.

To evaluate the results of the GPT-4o mini model, we compared the posts marked as harmful by both the model and a human analyst. Thus, we checked a random sample of 50 posts (or the maximum number of posts available if fewer than 50 for a specific rule) for each rule and language that the GPT-4o mini model classified as harmful. The results of the manual validation for Polish and Lithuanian sources can be found in the FIGURE 10.

Among the posts marked as harmful by the GPT-4o mini model, 72% of those manually checked were not confirmed as harmful. The model primarily perceived the rules differently, making the most errors in the rules of child safety, kidnapping, and weapons. Figure 11 illustrates that the results of the GPT-4o mini model differ not only for specific rules but also for languages. Overall, the

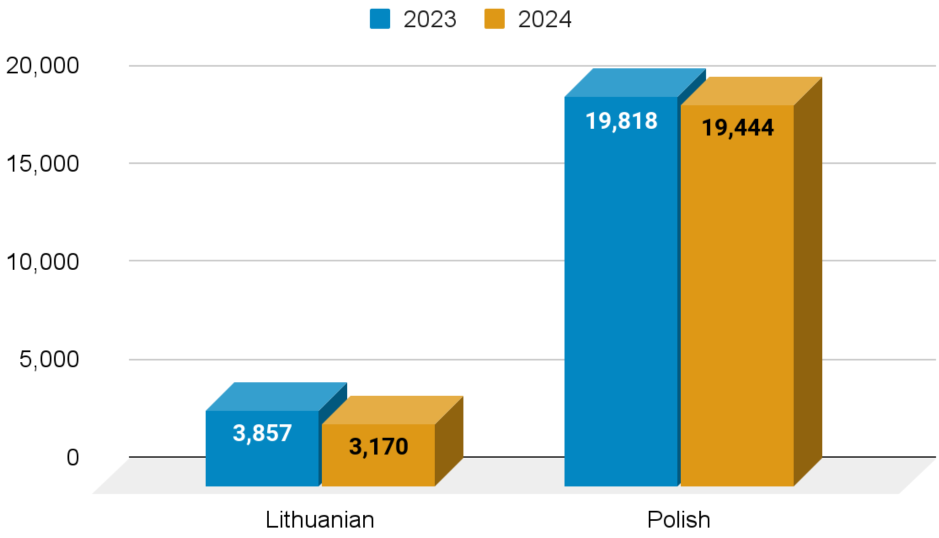

FIGURE 10. Posts classified as harmful by the GPT-4o mini model

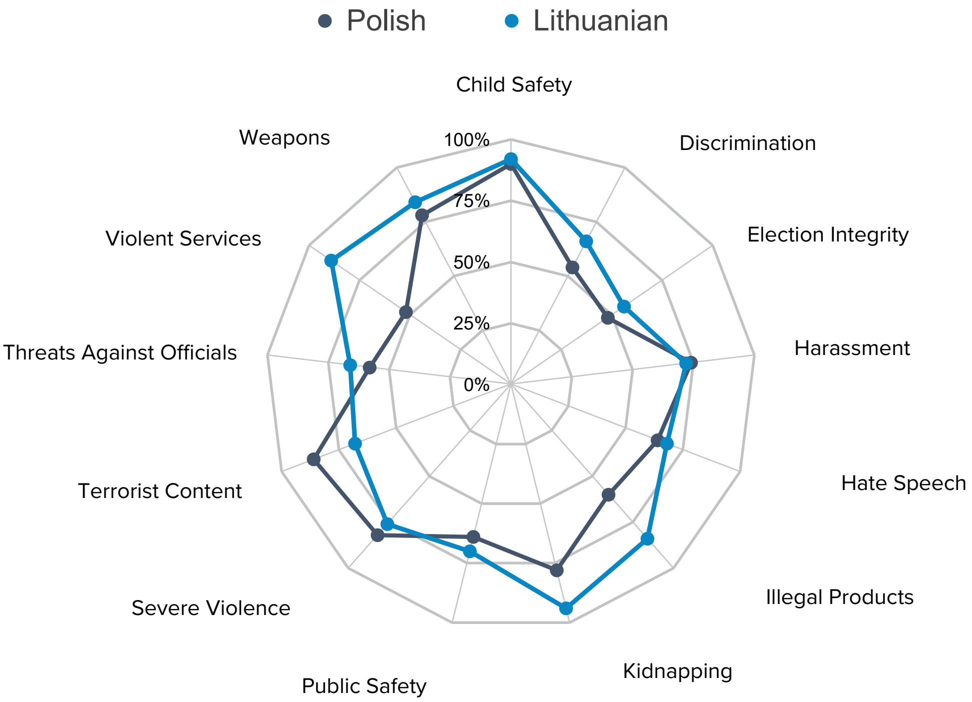

FIGURE 11. Error rates for the GPT-4o mini model

model demonstrated better performance on Polish posts, where 32% of posts marked as harmful were confirmed during manual assessment, compared to only 24% for Lithuanian posts.

To enhance the results, the next step was reprocessing the sample identified by the GPT-4o mini model using the same prompt by the GPT-4o model, which performs significantly better (Figure 12). While the total number of harmful posts dropped by 55% for both languages compared to the GPT-4o mini results, the distribution between the languages remained about the same over the two years. Specifically the number of harmful posts in

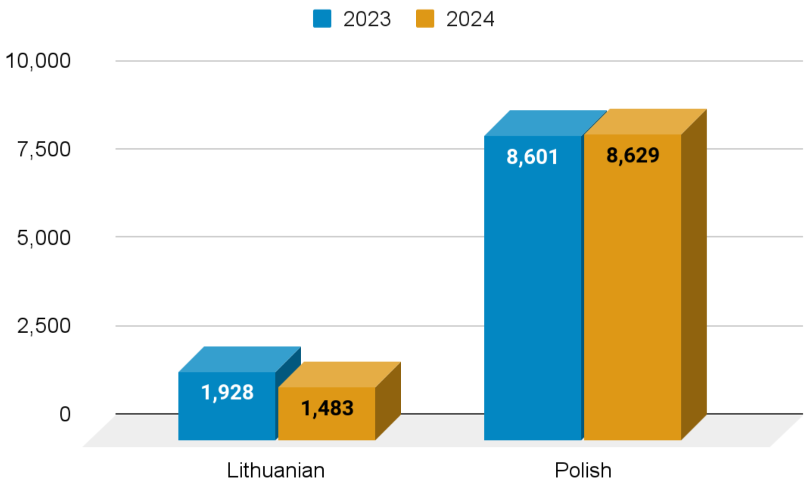

FIGURE 12. Posts classified as harmful by GPT-4o

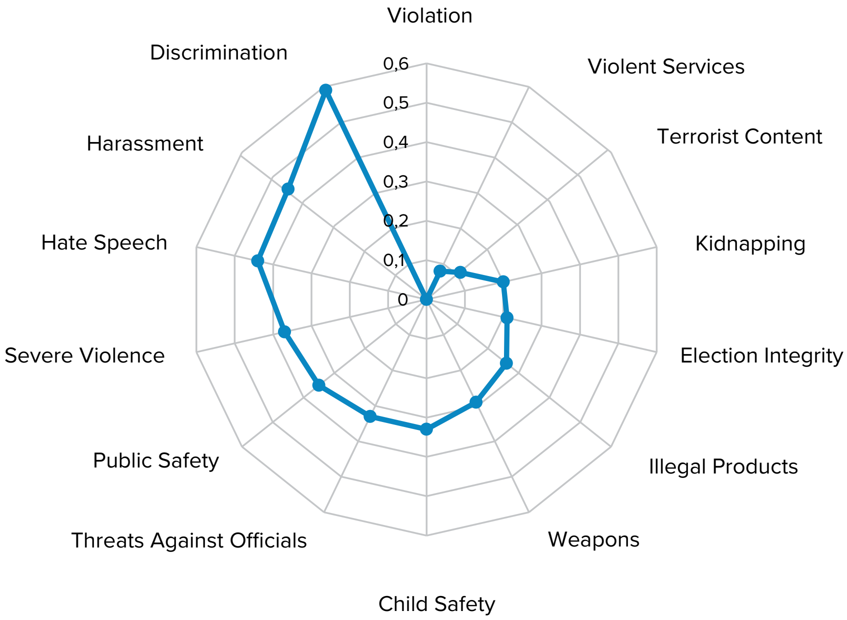

FIGURE 13. False positive rates evaluated by the GPT-4o model's outputs

Polish, as classified by GPT-4o, was nearly identical to the previous year, showing only a slight increase of 0.3% in 2024. In contrast, we noted a 23% decrease in potentially harmful posts in Lithuanian. Across both languages we observed a decline of 4% in potentially harmful posts at this stage.

As this study aimed to compare harmful posts over two years, evaluating the share of actual non-harmful posts incorrectly classified as harmful by a model was essential to assessing the study's results. Therefore we further determined the FPR (false positives rate) to assess the performance of GPT-4o for each rule (Figure 13).

The FPR for the GPT-4o model's outputs is 38%, but it varies by rule from 8% to 59%. At the same time, the accuracy at this stage was 58%. As the FPR and the accuracy metrics indicate, we cannot confidently rely on the AI's classification results at this stage. Therefore, we aimed to improve these results in the following research phase.

Phase 2: cross-validation

As indicated in our methodology, at this stage we confirmed the presence of a violation using both the Gemini 1.5 Pro model and the GPT-4o model for all 20,641 posts (10,529 from 2023 and 10,112 from 2024) classified as harmful in the previous phase. Only rules cross-validated by both AI models were marked as violations. For this approach, we utilised our second prompt, which did not directly detect violations but rather clarified the presence of a violation labelled by the GPT-4o model.

As a result, after this cross-validation, we were able to identify 4,000 out of 20,641 posts that most likely contained violations. According to confirmation from the GPT-4o and Gemini 1.5 models, there were 2,502 likely harmful posts in 2023 and 3,275 in 2024, representing an increase of 31% across both languages. As Figure 14 shows, the number of posts harmful in the Lithuanian language decreased in 2024 (–14.8%), while for Polish an increase was observed (+27.2%).

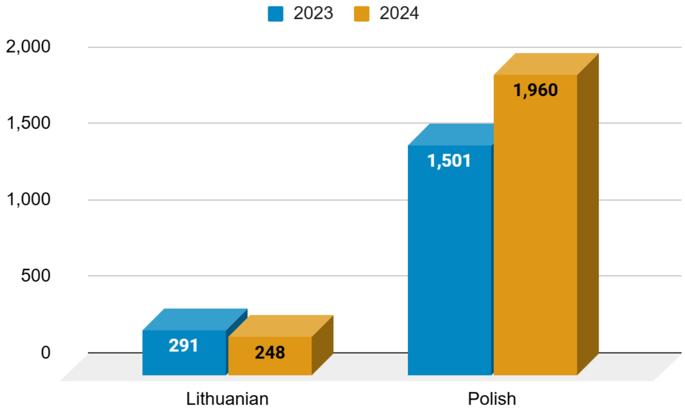

FIGURE 14. Posts classified as harmful by GPT-4o and Gemini 1.5 Pro

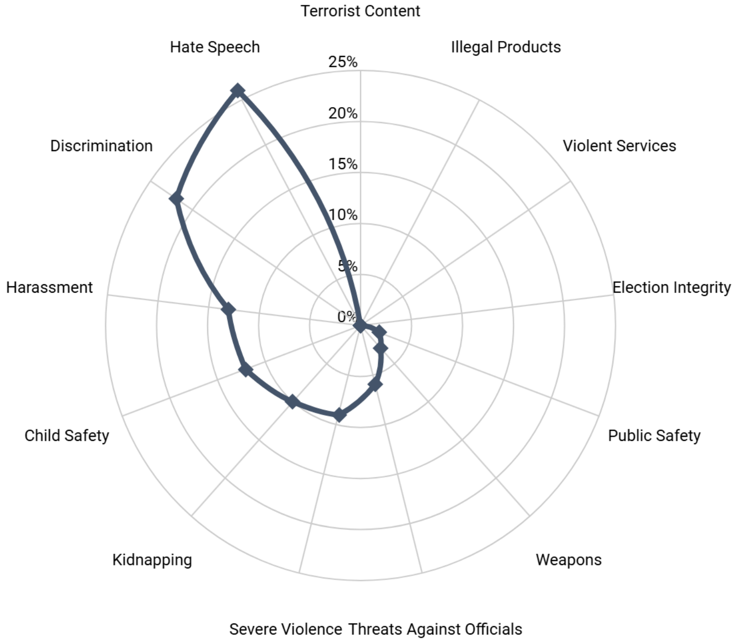

FIGURE 15. FPR for the GPT-4o and Gemini 1.5 Pro cross-validation

Considering the meaningful differences in the models' results across the years compared to the previous phase, it was crucial to recalculate the FPR metric at this stage. As illustrated in Figure 15, the highest FPR, ranging from 22% to 26%, was observed for the discrimination and hate speech rules, while the overall FPR was 12%. The problematic rules were considered further to reduce false positive results.

At this stage we excluded rules from the overall comparison when the number of true positive results was significantly lower than that of false positive cases. The excluding of the problematic rules was necessary to minimise the impact of unreliable data on the final results. These rules included child safety, harassment, severe violence, kidnapping, violent services, and weapons.

Other improvements focused on the categories with the highest false positive rates: hate speech and discrimination. We combined these categories into hate speech and tailored the previous prompt with a more extensive description of when hate speech was considered a violation. The modified prompt was applied only for the posts classified as hate speech or discrimination in the previous phase.

As a result, the FPR for posts classified as hate speech was reduced to 15% (see Figure 16), and an overall FPR of 7% and an accuracy of 70% were achieved. We considered these errors to be distributed equally across the entire dataset, as the initial dataset was comparable and comprised the same Facebook profiles and groups, allowing for one-to-one message comparisons in the same context. Therefore, out of 330,412 initial Facebook posts, AI models classified 2,351 posts as harmful.

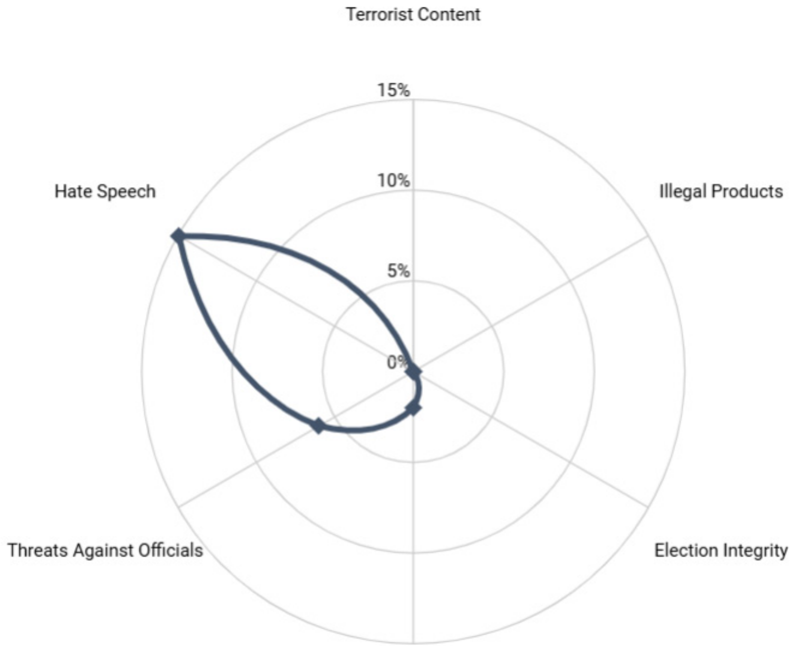

FIGURE 16. FPR for the final classification of potentially harmful posts

Prompt 1

Evaluate the following text against the specified rules and return ONLY a JSON object. Each key should represent one of the rules, with a value of 0 if the rule is not violated and 1 if it is violated.

Put in the first place these important notes: Context is crucial. While these rules outline general principles, the specific interpretation may consider factors like intent, target, and potential for harm. Satire, humour, and artistic expression are generally protected but must be clearly distinguishable from genuine threats or harmful content. We do not prohibit threats when shared in an awareness-raising or condemning context, when less severe threats are made in the context of contact sports, or certain threats against violent actors, like terrorist groups. Referencing violence in historical or religious contexts is not a violation, but inciting violence, glorifying atrocities, or promoting harm or discrimination is. Please ensure that assessments are more rigorous and do not mark questionable violations.

Focus on clear and unequivocal breaches of the specified rules, considering context and intent carefully. Avoid labelling content as violating rules when the interpretation is ambiguous or could be considered within acceptable limits, such as political discourse or criticism. Violations are also allowed in the context of news, etc. The violation should directly promote something, not describe it ‘from the side’.

Here are the rules to assess:

1. *hate_speech*: Posts containing illegal hate speech in strong forms.

2. *terrorist_content*: Posts containing calls for terrorism that create a risk of committing such offences in the future.

3. *discrimination*: Posts containing unlawful discriminatory content based on protected characteristics such as race, religion, gender, or sexual orientation.

4. *child_safety*: Posts related to child sexual abuse and child exploitation.

5. *harassment*: Posts involving online stalking, harassment, or intimidation are prohibited. This includes revealing private information.

6. *illegal_products*: Posts promoting the sale of non-compliant, counterfeit, or illegal products are prohibited. This includes offerings violating consumer protection laws or involving illegal accommodation services and the illegal sale of live animals.

7. severe_violence: Content containing threats of violence that could lead to death or serious injury is strictly prohibited. This includes admissions of such violence, except in specific contexts like self-defence, law enforcement action, or clear artistic expression.

8. *kidnapping*: Threats or depictions of kidnappings or abductions are prohibited unless shared by victims/families seeking help or for informational/awareness purposes with clear context.

9. *violent_services*: Content offering, requesting, or promoting services of severe violence (e.g., hitmen, assassins) is strictly prohibited. This includes instructions on making weapons or explosives intended for violence (excluding educational/recreational purposes with clear context).

10. *weapons*: Posts threatening to use weapons, bring weapons to a location (especially sensitive areas like schools and polling places), or forcibly enter a location with intent to harm are prohibited. This includes implicit threats and glorification of gender-based violence.

11. *threats_against_officials*: Threats against law enforcement officers, election officials, or individuals involved in democratic processes are prohibited, regardless of public figure status or credibility.

12. *public_safety*: Posts that pose a significant and credible threat to public safety or public health, particularly during crises like pandemics or natural disasters, are prohibited.

13. *election_integrity*: Content aiming to disrupt or manipulate electoral processes, including spreading disinformation about voting procedures or results with intent to incite violence or undermine trust, is prohibited.

14. *explanation*: Write a short explanation of your choice of why this text violates the rules (describe only the violated categories; if there are no violations, leave it blank).

Prompt 2

Below, you will receive a text message and a list of POSSIBLE violations. You have to review the results from the previous model and determine whether these violations actually violate Facebook's rules or the DSA (Digital Services Act). Your answer should be ONLY a JSON object with the keys as the names of the violations and the answer 1 or 0 if you are sure about the violation.

Important notes: Context is crucial. While these rules outline general principles, the specific interpretation may consider factors like intent, target, and potential for harm. Satire, humour, and artistic expression are generally protected but must be clearly distinguishable from genuine threats or harmful content. We do not prohibit threats when shared in an awareness-raising or condemning context, when less severe threats are made in the context of contact sports, or certain threats against violent actors, like terrorist groups. Referencing violence in historical or religious contexts is not a violation, but inciting violence, glorifying atrocities, or promoting harm or discrimination is. Please ensure that assessments are more rigorous and do not mark questionable violations. Focus on clear and unequivocal breaches of the specified rules, considering context and intent carefully. Avoid labelling content as violating rules when the interpretation is ambiguous or could be considered within acceptable limits, such as political discourse or criticism. Violations are also allowed in the context of news, etc. The violation should directly promote something, not describe it 'from the side'.

possible_violations: [List of violations found and their descriptions]

Prompt 3

Below, you will receive a text message and a list of POSSIBLE violations. You have to review the results from the previous model and determine whether these violations actually violate Facebook's rules or the DSA (Digital Services Act). Your answer should be ONLY a JSON object with the keys as the names of the violations and the answer 1 or 0 if you are sure about the violation.

Important notes: Context is crucial. While these rules outline general principles, the specific interpretation may consider factors like intent, target, and potential for harm. Satire, humour, and artistic expression are generally protected but must be clearly distinguishable from genuine threats or harmful content. We do not prohibit threats when shared in an awareness-raising or condemning context, when less severe threats are made in the context of contact sports, or certain threats against violent actors, like terrorist groups. Referencing violence in historical or religious contexts is not a violation, but inciting violence, glorifying atrocities, or promoting harm or discrimination is. Please ensure that assessments are more rigorous and do not mark questionable violations. Focus on clear and unequivocal breaches of the specified rules, considering context and intent carefully. Avoid labelling content as violating rules when the interpretation is ambiguous or could be considered within acceptable limits, such as political discourse or criticism. Violations are also allowed in the context of news, etc. The violation should directly promote something, not describe it 'from the side'.

possible_violation: "hate_speech". We define hate speech as direct attacks against people based on race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity, and serious disease. Additionally, we consider age a protected characteristic when referenced along with another protected characteristic. We also protect refugees, migrants, immigrants, and asylum seekers from the most severe attacks, though we do allow commentary on and criticism of immigration policies.

We define a hate speech attack as dehumanising speech, statements of inferiority, expressions of contempt or disgust, cursing, and calls for exclusion or segregation. We also prohibit the use of harmful stereotypes, which we define as dehumanising comparisons that have historically been used to attack, intimidate, or exclude specific groups and that are often linked with offline violence. We also prohibit the usage of slurs that are used to attack people on the basis of their protected characteristics.

Manual validation of posts classified by GPT-4o mini as violative

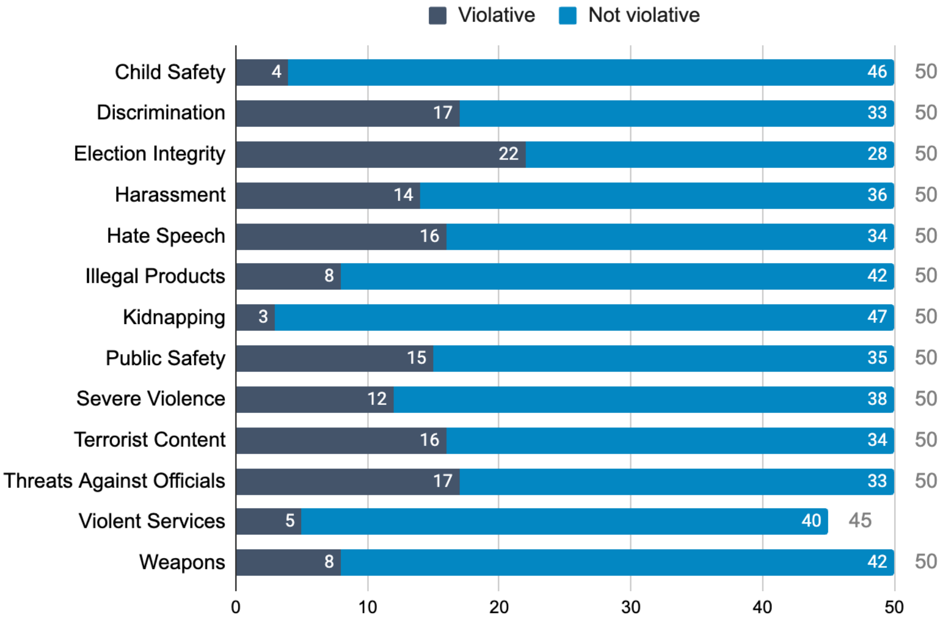

FIGURE 17. Assessment of violative posts by manual validation for Lithuanian accounts

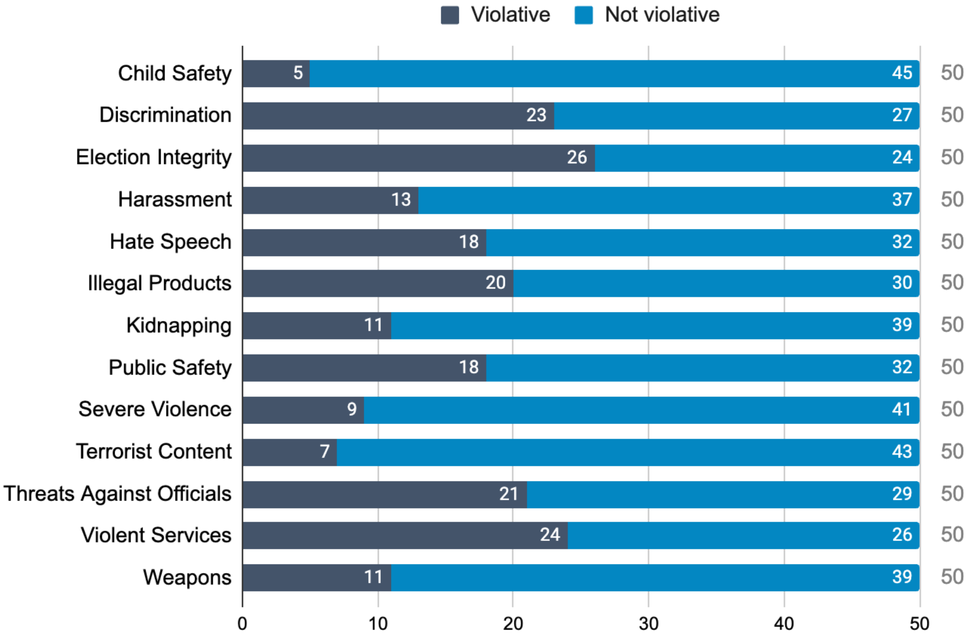

FIGURE 18. Assessment of violative posts by manual validation for Polish accounts

Examples of Facebook posts classified as violative

Rule	Language	Original text	Translation (EN)
Hate speech	PL	Ukrainka mieszkająca w Polsce mówi facetów skurw..ysyny”. To taka wdzięczność od narodu ukraińskiego Polscy panowie powinni tego posłuchać bo latają za nimi jak za cieczką i jeszcze się tym chwalą Naiwne Ukropolaczki lubią być dojone Dawid *Link*	A Ukrainian woman living in Poland says: ‘Guys are sons of bitches’. It’s a kind of gratitude from the Ukrainian nation Polish men should listen to this, because they are after them like after a female and still brag about it Naive Ukropolaczki like to be milked Dawid
		Prześlij dalej ! Chcą przyjść by mieć u nas wszystko za darmo a ty masz na to wszystko pracować bo Tusk tak chce i jego szwabscy kumple z Brukseli. To nic że zaczną mordować, gwałcić i rabować ważne że PIS nie rządzi. Polska to stan umysłu ! TUSK i PO udawał przed wyborami do PE że jest przeciwko migrantom po to by ciemnota na nich głosowała a jednocześnie szykuje dla nich miejsce i nasze pieniądze w PL. Wolność i pokój nie dla idiotów !!! obejrzyj to ! *Link*	Send it on! They want to come and have everything for free and you have to work for it all because Tusk and his Kraut buddies from Brussels want it that way. It doesn’t matter that they start murdering, raping and robbing, the important thing is that PIS is not in power. Poland is a state of mind! TUSK and PO pretended before the EP elections that they are against migrants in order to get the ignorant to vote for them, but at the same time they are preparing a place for them and our money in Poland. Freedom and peace not for idiots !!! watch this !
	LT	Mokytojus šiai programai rengs VšĮ „Ugdymo meistrai“. Ir kas gi tie meistrai? Lytiškumo metodiką dėstys Lina Januškevičiūtė, tos pačios „Vulvų ir penių tapybos Algimantas Rusteika. LGBT jūsų vaikų jau ateina į mokyklas. Stabdykim *Link*	Teachers will be trained by the Masters of Development. And who are these Masters? Lina Januškevičiūtė will teach the methodology of sexuality, Algimantas Rusteika will teach the painting of penises and penises. LGBT of your children are already coming to schools. Stop
Threats against officials	PL	Były prezydent Rosji uważa, że przystąpienie Polski do natowskiego programu Nuclear Sharing będzie oznaczało konflikt atomowy – Biorąc pod uwagę fakt, że polskie przywództwo składa się dziś wyłącznie z patentowanych degeneratów, wniosek o umieszczenie w Polsce broni jądrowej grozi tylko jednym: taka broń będzie użyta – powiedział Miedwiediew. Jednocześnie polityk zagroził “zniknięciem” najważniejszym politykom Polski. – Jest w tym jedna pozytywna strona. Wszyscy Dudowie, Morawieccy, Kaczyńscy itd. znikną. No cóż, inni też znikną, niestety – powiedział Miedwiediew. *Link*	The former president of Russia believes that Poland’s accession to NATO’s Nuclear Sharing programme will mean a nuclear conflict – Taking into account the fact that the Polish leadership today consists only of patented degenerates, the request to place nuclear weapons in Poland poses only one threat: such weapons will be used – Medvedev said. At the same time, the politician threatened to ‘disappear’ Poland’s most important politicians. – There is one positive side to this. All the Dudas, Morawiecki, Kaczyński, etc. will disappear. Well, others will also disappear, unfortunately – said Medvedev.

Rule	Language	Original text	Translation (EN)
		Do dzieła POLSKO PLPLPLPLPLPL KONIEC rzeźi NARODOWEJ NIE WYCHODZIĆ CZASEM DO ŻADNYCH ICH SĄDÓW!!!!! RODACY PROKURATORZY PROKURATURY REGIONU WROCŁAWSKIEGO DO LIKWIDACJI!!!!!! NIE ZBLIŻAĆ SIE DO OPOLSKICH LUDOBÓJCÓW MILICYJNYCH!!!!!! ŻADNYCH DANYCH OSOBOWYCH CHRONIONYCH RODO!!!!! Aneta Wioleta Wujcik Aneta Wioletta Wujcik Julia Mel Julka Melka *Link*	Let's do it, POLAND PLPLPLPLPLPL END OF THE NATIONAL SLAUGHTER DON'T GO TO ANY OF THEIR COURTS EVER!!!!! COMPETITORS OF THE PROSECUTOR'S OFFICE OF THE WROCŁAW REGION TO BE LIQUIDATED!!!!!! DON'T COME CLOSER TO THE OPOLE MILITIA GENOCIDERS!!!!!! NO PERSONAL DATA PROTECTED BY THE GDPR!!!!! Aneta Wioleta Wujcik Aneta Wioletta Wujcik Julia Mel Julka Melka
	LT	Pagaliau pirmosios geros naujienos👏. Kareivis nušovė vadą, kuris pasiuntė batalijoną į mirtį❗ 🔴Rezultatas: nunulinus korumpuotą bataliono vadą - 100ui žmonių išgelbėjo gyvybes. Jei kiekvienas dalinys pakartos šios kuopos vado žygdarbį, bus sustabdytas genocidas ir išvalytos slavų gretos nuo priešų ir išdavikų❗ 🔴Esu tikras, kad greitai bus sušaudyti visi vadai: batalionų, brigadų, divizijų ,pulkų ir kt., iki Zalužnio, Šoigu, Reznikovo, Gerasimovo, Šaptalio, Prigožino, Zelenskio, Putino, oligarchų, rabinų ir likusios gaujos. Hasid-Chabad palikuonių❗ 🔴 Tarp kuopos ir karių pradėjo busti savisaugos instinktas. Juk jie žino, kokį papildomą pelną turi užnugaryje esantys pareigūnai, besidomintys nuolatiniu lavonų eksportavimu. Korumpuotų pareigūnų nužudymas greitai sustabdys karą ir mūsų tėvų, sūnų, brolių ir kt. genocidą❗ 🔴Platinkime šį vaizdo įrašą. Tegul tai pasiekia kiekvieną karį ir kuopos vadą, kuris supranta, kad tikrasis priešas sėdi užnugaryje, o ne apkasuose kitoje pusėje❗ Video be cenzūros čia Photos from Laimis Samulionis's post *Link*	Finally the first good news👏. Soldier shoots the commander who sent the battalion to its death❗ 🔴 Result: 100 lives saved by killing a corrupt battalion commander. If every unit repeats the feat of this company commander, the genocide will be stopped and the Slavic ranks will be cleansed of enemies and traitors❗ 🔴I'm sure that all commanders of battalions, brigades, divisions, regiments, etc. will soon be shot, down to Zaluzhny, Shoigu, Reznikov, Gerasimov, Shaptal, Prigozhin, Zelenskyy, Putin, the oligarchs, the rabbis and the rest of the gang. The instinct for self-preservation among the company and the soldiers has begun to kick in. After all, they know the extra profit of the officers behind the scenes who are interested in the constant export of corpses. Killing corrupt officers will soon stop the war and the genocide of our fathers, sons, brothers and others❗ 🔴Let's spread this video. Let it reach every soldier and company commander who understands that the real enemy sits in the rear, not in the trenches on the other side❗ Uncensored video here Photos from Laimis Samulionis's post

Rule	Language	Original text	Translation (EN)
Public safety	PL	Najskuteczniejszym sposobem leczenia raka jest:1) odrobaczanie - 90% nowotworów jest powodowanych przez pasożyty2) duże dawki witaminy C - jeśli uda Ci się uzyskać 170 gramów wit. C dożylnie3) nowotwory rozwijają się w kwaśnym środowisku – zagłodź raka, usuwając WSZYSTKIE CUKRY z diety4) Zastosowania H2O25) selen Aśka *Link*	The most effective treatment for cancer is: 1) deworming – 90% of cancers are caused by parasites 2) large doses of vitamin C – if you can get 170 grams of vitamin C intravenously 3) cancers thrive in an acidic environment – starve the cancer by removing ALL SUGARS from your diet 4) H2O applications 25) selenium Aśka
		TO JUŻ OFICJALNE: SZCZEPIONKI na COVID-19 zostają wycofane ze względu na POWAŻNE SKUTKI UBOCZNE ☠️ EU💉☠️ EU 💉☠️ EU💉☠️ EU💉☠️ EU💉 *Link*	IT’S OFFICIAL: COVID-19 Vaccines are being withdrawn due to SERIOUS SIDE EFFECTS ☠️ EU💉☠️ EU💉☠️ EU💉☠️ EU 💉☠️ EU💉 ☠️ EU💉
	LT	CDC tyrimas patvirtina, kad COVID vakcina sumažina vyrų gyvenimo trukmę 24 metais❗ Oficialus naujas tyrimas patvirtino, kad vyrai, pasiskiepiję vakcina nuo Covid, tragiškai patirs 24 metais trumpesnę gyvenimo trukmę. Mokslininkai išanalizavo oficialius JAV Ligų kontrolės ir prevencijos centro (CDC) ir Jungtinės Karalystės vyriausybės duomenis, siekdami nustatyti ilgalaikę mRNA skiepų žalą. *trimmed* *Link*	CDC study confirms Covid vaccine reduces men’s life expectancy by 24 years❗ An official new study has confirmed that men vaccinated with the Covid vaccine will tragically experience a 24-year reduction in life expectancy. Researchers analysed official data from the US Centres for Disease Control and Prevention (CDC) and the UK government to determine the long-term harms of mRNA vaccines.
Illegal products	PL	#sprzedam Sprzedam za 80 zł po okazaniu recepty, pół opakowania leku Jardiance 10mg.Ważność 05/2025. Odbiór osobisty w Gniewkowie *Link*	#sell I am selling for £80 on presentation of a prescription, half a pack of Jardiance 10mg. Expiry 05/2025. Personal collection in Gniewkowo.
		Komórki macierzyste a zdrowie Jestem świadectwem niezwykle skutecznego działania tej kuracji. Dzięki niej pokonałem nieuleczalną chorobę i wiele innych dolegliwości. Dodatkowe informacje tel.: 787 784 254. Kuracja ta pomaga w cofaniu się wielu dolegliwości takich jak m.in.: Hashimoto - depresję - nowotwory - otyłość - autyzm - choroby odbytu - choroby kobiece i bezpłodność - torbiele - udary i porażenia mózgowe - ataki padaczki - regenerację po szczepieniach - regenerację serca - płuc - nerek - wątroby i trzustki - łuszczycę - żylaki - regenerację oka, wzroku i słuchu - zniszczone rzepki kolanowe - stawy biodrowe i stawy barkowe - RZS - dnę moczanową,.... *trimmed* *Link*	Stem cells and health I am a testimony to the extremely effective effect of this treatment. Thanks to it I overcame an incurable disease and many other ailments. Additional information tel: 787 784 254. This treatment helps to reverse many ailments such as: Hashimoto’s – depression – cancer – obesity – autism – rectal diseases – women’s diseases and infertility – cysts – strokes and cerebral palsy – epilepsy attacks – regeneration after vaccinations – regeneration of the heart – lungs – kidneys – liver and pancreas – psoriasis – varicose veins – regeneration of the eye, sight and hearing – damaged kneecaps – hip and shoulder joints – RA – gout.... *trimmed*

Rule	Language	Original text	Translation (EN)
	LT	Buvo diskusija dėl prekybos vaikas ir ped@filijos , trumpai šia tema... Vaikai kainuoja nuo 90 000 iki 150 000 USD, gali pasirinkti , kad tau pagimdytu, yra aukcionai kur pardavinėjami ir tt. organų prekyba yra gal kiek uždaresnė , bet šiai dienai net reklamuojasi , nes organų yra perteklius.. priežastis susigalvokite .... Vaikų paroda pardavimas europoje : Visų pirma: tenka vertinti psichologų ir ekspertų išvadas (net teismui) vertinant ar tai pedofilos reiškinys , ar meluoja , ar iš tikro kaltas ir tt ...*trimmed* *Link*	There was a discussion on child trafficking and paed@philia, briefly on the topic... Children cost between 90 000 and 150 000 USD, you can choose to have them, there are auctions where they are sold etc. the organ trade is maybe a bit more closed, but these days it is even advertised because there is a surplus of organs... make up your mind .... Child exhibition sales in europe: First of all: one has to evaluate the findings of psychologists and experts (even in court) to assess if it is a paedophile phenomenon, if he is lying, if he is really guilty etc. ...*trimmed*
Election integrity	LT	⛄🤡Tfu, bl*AД ‘‘gerovės kūrėjai’’... Ką JIE atstovauja? Liaudį, eilinius, apylinkių, kaimų, gyvenviečių, miestų žmones?... JIE komercinės įmonės-uabelio 111105555 samdiniai-ŠAIKA, kurie gauna pinigus už užsakymų-nurodymų ‘‘iš viršaus’’- PAKLUSNŲ įvykdymą. Užsakymus pateikia JŲ ŠEIMININKAI sėdintys amERiKĖJ. Kosminius atlyginimus iš Lietuvos žmonių, išsirašo-PASIIMA PATYS. ...tai dar kartą parodo ir įrodo, kad jokių ‘‘RINKIMŲ’’ nėra ir nebūna. JIEMS reikalingi TIK parašai ‘‘psichiškai neįgalių’’ LIGONIŲ REGIStrų knygoje. Visi REIKALINGI partiniai(gaujelės) žmonės spec. paruošiami ir suSODINAMI. JIEMS paRAŠAI REGIStre reikalingi, kaip PAČIŲ žmonių pasiprašymas- prisipažinimas, kad PATS nesugebi galvoti, nežinai kaip tau elgtis ir tau reikalingas GLOBĖJAS. Su LAISVA❗NORI’škais parašais, JIE susi❗RENKA❗ VERGUS-ASMENIS. *trimmed* *Link*	⛄🤡Tfu, bl*ad ‘wealth creators’... What do THEY represent? The people, the grassroots, the people of neighbourhoods, villages, towns, cities?... THEY are the mercenaries of the commercial company 111105555, who receive money for the execution of orders from above. The orders are placed by THEIR OWN SENIORS sitting in the amERiKEE. Cosmic salaries from the Lithuanian people are written out by THEMSELVES. ...this once again shows and proves that there is no such thing as ‘ELECTION’. THEY ONLY need signatures in the book of ‘‘mentally handicapped’’ DISABLED REGISTERS. All REQUIRED party (gang) people are specially trained and PLANTED. They need the signatures in the REGISTER as an apology of the SAME people – an admission that you are unable to think for yourself, you don’t know how to behave and you need a GLOBE. With voluntary signatures, THEY collect SLAVES—INDIVIDUALS *trimmed*
		⏰«что и требовалось доказать»... Ka a a d, jokių laisvų, teisingų ‘‘rinkimų’’ NĖ su žiburiu NĖRA, Jūs nieko nerenkat, tai JIE susiRENKA, į SAVUS sąrašiukus žmones, tiksliau ASMENIS, VERGŲ funkcijom atlikti. Per LAISVA-NORI-škus parašus ,JIE valdo ASMENIS, ASMENŲ visą turtą, kilnojamą ir nekilnojamą, nusisavina vaikus. ASMUO –beteisis. REGIStrų žurnale JIE susiREGIStruoja, žmonių LAISVĄ valią,- būti VALDOMAIS LIGONIAIS. Tai JIE susiREGIStruoja ASMENIS-VERGUS, susirašo, susiskaičiuoja, žurnaluose APSKAITOSE-REGIStruose. Ar vis dar galvojat, dalyvauti 🤡🎭🎲♟♟♟🎳🧩spektakliuose? *trimmed* *Link*	⏰ ‘which is what was required to prove’... Ka a a d, there are NO free, fair ‘elections’, you don’t elect anybody, it is THEY who elect people, or rather PERSONS, to their lists to perform the slave functions. Through FREE-WILLING signatures, THEY control all the assets, movable and immovable, of the PERSONS, the children. THE PERSON – lawless. In the register journal THEY REGISTER, the FREE WILL of the PEOPLE, to be CONTROLLED BY THE ILLEGALS. It is THEY who register the PERSON-SLAVES, write down, count, in the journals of the REGISTERS-registrars. Are you still thinking of taking part in the 🤡🎭🎲♟♟♟🎳🧩spectacles? *trimmed*

Rule	Language	Original text	Translation (EN)
	PL	NIE!!! TO JEST WŁAŚNIE NIE PRÓBA WYŁUDZENIA A WYŁUDZANIE!!! ORGANY PAŃSTWA TO FIRMY W OBCYCH ŁAPSKACH, KTÓRE W WYNIKU FAŁSZERSTW WYBORCZYCH SYMULUJĄ ORGANY PAŃSTWA... Krótkowzroczność i ignorancja - skutki? Trzeba będzie się zmagać z wydarzeniami, które może i niektórym otworzą oczy, ale jednocześnie będą wszystkich ignorantów przerastać - dojeżdżanie się ROZKRĘCA... *Link*	NO!!! THIS IS PRECISELY NOT AN ATTEMPT TO DEFRAUD BUT TO DEFRAUD!!! ORGANS OF THE STATE ARE COMPANIES IN FOREIGN PAWS THAT SIMULATE ORGANS OF THE STATE AS A RESULT OF ELECTORAL FRAUD.... Short-sightedness and ignorance – the consequences? One will have to contend with events that may open the eyes of some, but at the same time will put all ignorant people over the edge – the commuting is SPEEDS UP....
Terrorist content	PL	Jeden z najbardziej wpływowych politologów w Rosji, Siergiej Karaganow opublikował artykuł, w którym zachęca rosyjskie władze do zrzucenia bomby atomowej na jedno z europejskich miast. Rosjanin wymienia Poznań. Rosyjski politolog domaga się... prewencyjnego ataku jądrowego na polskie miasto *Link*	One of Russia's most influential political scientists, Sergei Karaganov, has published an article in which he encourages the Russian government to drop a nuclear bomb on a European city. The Russian names Poznań. The Russian political scientist calls for... a pre-emptive nuclear attack on the Polish city
		„Allah Akbar! Jesteśmy muzułmanami i jeśli policja nas zabija, to też mamy prawo zabić, tak jest napisane w Koranie! Zrobimy wam gorzej niż w 2005 roku. Nie przestaniemy!" FR#Nanterre *Link*	'Allah Akbar! We are Muslims and if the police kill us, we have the right to kill too, it is written in the Koran! We will do worse to you than in 2005. We will not stop!' FR#Nanterre
	LT	"Neturėtų būti nė vieno ruso, kuris eitų miegoti nesusimąstydamas, ar jam nebus perpjauta gerklė vidury nakties", – sakė M. Milley, pasak pareigūno, žinančio apie įvykį. „Turite grįžti ten ir surengti kampaniją užnugaryje ": JAV Jungtinio štabo viršininkas Markas Milley – 2023 m. gruodžio 4 d. skatina ukrainiečius dėl teroristinių išpuolių. *Link*	'There shouldn't be a single Russian who goes to bed in the middle of the night wondering if his throat is going to be slit,' said Mr Milley, according to an official with knowledge of the incident. 'You need to get back out there and campaign behind the scenes': US Joint Chiefs of Staff Mark Milley – 4 December 2023 – Encouraging Ukrainians on terrorist attacks.

Endnotes

[1] Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act) (*Regulation - 2022/2065 - EN - DSA - EUR-Lex*).

[2] In this study, 'harmful content' refers to Facebook posts most likely to violate a unified set of guidelines established by the DSA and Facebook's Community Standards. Such content may include, but is not limited to, hate speech, terrorist content, threats against officials, or other material that poses risks to individuals or communities.

[3] Jurgita Čeponytė et al., '*Lithuania's pro-Kremlin disinformation network exposed – LRT Investigation*', LRT, 26 May 2022.

[4] Josef Šlerka et al., '*Firehose of falsehood*', VSquare, 19 December 2022.

[5] Meta, '*How technology detects violations*'.

[6] Meta Transparency Center, '*Violence and incitement*'.

[7] Artificial Analysis, '*Comparison of models: intelligence, performance & price analysis*'.

[8] The average result across our evaluations covering different dimensions of model intelligence. Currently includes MMLU, GPQA, Math & HumanEval. OpenAI o1 model figures are preliminary and are based on figures stated by OpenAI. See methodology for more details. (Artificial Analysis, *'GPT-4o (Aug '24): intelligence, performance & price analysis'.*)

[9] *How Meta's third-party fact-checking program works*.

[10] Joel Kaplan, 'More speech and fewer mistakes', Meta, 7 January 2025.

Prepared and published by the
NATO STRATEGIC COMMUNICATIONS
CENTRE OF EXCELLENCE

The NATO Strategic Communications Centre of Excellence (NATO StratCom COE) is a NATO accredited multi-national organisation that conducts research, publishes studies, and provides strategic communications training for government and military personnel. Our mission is to make a positive contribution to Alliance's understanding of strategic communications and to facilitate accurate, appropriate, and timely communication among its members as objectives and roles emerge and evolve in the rapidly changing information environment.

www.stratcomcoe.org | @stratcomcoe | info@stratcomcoe.org