# **Complaints about content moderation procedures on social media:** what users think

iris
INSTITUTO
DE REFERÊNCIA
EM INTERNET
E SOCIEDADE

# **Complaints about content moderation procedures on social media:** what users think

**AUTHORSHIP**
Fernanda dos Santos Rodrigues Silva
Júlia Maria Caldeira Gertrudes
Luiza Correa de Magalhães Dutra
Rafaela Ferreira Gonçalves da Silva

**REVIEW**
Paloma Rocillo
Paulo Rená da Silva Santarém

**TRANSLATION**
Fernanda dos Santos Rodrigues Silva
Luiza Correa de Magalhães Dutra
Paulo Rená da Silva Santarém

**GRAPHIC DESIGN, COVER, LAYOUT, IMAGE MANIPULATION AND FINALIZATION**
Felipe Duarte
Images: Freepik

**EDITORIAL PRODUCTION**
IRIS - Instituto de Referência em Internet e Soceidade

**HOW TO REFERENCE IN ABNT**
SILVA, Fernanda dos Santos Rodrigues; GERTRUDES, Júlia Maria Caldeira; DUTRA, Luiza Correa de Magalhães; SILVA, Rafaela Ferreira Gonçalves da. **Complaints about the content moderation procedure on social networks**: what users think. Belo Horizonte: Internet and Society Reference Institute, 2024. Available at: https://bit.ly/3Nc1epW. Accessed at: dd mmm yyyy.

iris

INSTITUTE
FOR RESEARCH
ON INTERNET
AND SOCIETY

**DIRECTORS**

Ana Bárbara Gomes
Paloma Rocillo

**MEMBERS**

Felipe Duarte | Head of Communications
Fernanda Rodrigues | Head of Research and Researcher
Glenda Dantas | Researcher
Júlia Caldeira | Researcher
Júlia Tereza Koole | Research Trainee
Luísa Melo | Research Trainee
Luiza Correa de Magalhães Dutra | Researcher
Paulo Rená da Silva Santarém | Researcher
Thais Moreira | Communications Analyst
Wilson Guilherme | Researcher

irisbh.com

# Summary

# Executive Summary

- The present study sought to understand the main complaints from users regarding content moderation on social media, based on an analysis of complaints related to Facebook, Instagram, Twitter, TikTok, and YouTube submitted on the *Reclame Aqui* platform.

- Of the total 1,081 complaints collected, those unrelated to moderation were discarded (28.03%), those about hacking were excluded (23.03%), and cases where it was not possible to determine if they were related to moderation were also removed (7.4%), resulting in a final sample of 449 complaints that were effectively analyzed.

- Among the analyzed complaints, 54.34% are pertinent to the object of the research, specifically **complaints regarding the content moderation procedures related to the removal of posts and the suspension/blocking of accounts.** This highlights how the opacity in the moderation process contributes to user distrust, negatively impacting the public perception of platforms.

    - Considering only this universe, the most frequent complaints are, in descending order:
    - inadequate justification for moderation decisions (52.46%);
    - unanswered challenge to moderation decision (22.54%);
    - lack of notification or warning about moderation decision (9.02%);
    - lack of tools to challenge the moderation decision (7.38%);
    - inaccessible platform design in relation to moderation decision review mechanisms (4.51%);
    - others (4.1%).

- Among the analyzed complaints, the remaining 45.66% were related to content moderation, including procedures, but not specifically about our object. In this universe, the complaints were about:

    - generic complaints about platforms' content moderation (36.1%);
    - third-party content moderation requests (30.24%);
    - problems with monetization (10.24%);
    - problems with the recommendation or reach of the content (10.24%);
    - restriction of functionalities on the platform due to a moderation decision (10.24%);
    - problems generated by age limitations of the platform (1.95%);
    - ad moderation requests (1.46%).

- In light of this, **it is recommended** that digital platforms appropriately promote transparency and justification in moderation decisions. It is important to provide clear explanations regarding the reasons behind actions taken, such as content removal or account suspensions. Transparency helps ensure that decisions are not seen as arbitrary or unfair, and allow users to better understand the rules and how they are applied. Furthermore, justifying moderation decisions fosters a more equitable environment and can reduce users' feelings of censorship or injustice;

- **It is also recommended** that platforms enhance the effectiveness of their appeal processes and review their internal policies to effectively address users' needs and ensure a safe and trustworthy digital environment. Users must have easy access to effective mechanisms to challenge moderation decisions they consider unfair and these processes must be simple and efficient. This includes ensuring that appeals are reviewed fairly and impartially, with feedback on the results, leaving no requests unanswered, and using response automation sparingly.

# Presentation

IRIS is an independent and interdisciplinary research center dedicated to producing and communicating scientific knowledge on internet and society topics with the aim of promoting public policies that advance human rights in the digital area. For more than 5 years, IRIS has been specifically dedicated to understand the field about content moderation, with scientific research published on transparency,[1] due process[2] and strategies to combat disinformation.[3]

This document presents some of the initial findings from the research project **"Between Posts and Controversies: Conflict Resolution Strategies in Content Moderation on Digital Platforms."** Conducted independently with financial support from Google, the project aims to understand the issues related to the removal of legitimate content on digital platforms, particularly social media. This report specifically focuses on identifying the main complaints from users regarding moderation procedures, viewing this information as essential for enhancing platform operations and ensuring greater security for users.

The aim is to provide an evidence-based understanding of how controversies related to

[1] KURTZ, Lahis Pasquali; DO CARMO, Paloma Rocillo Rolim; VIEIRA, Victor Barbieri Rodrigues. **Transparência na moderação de conteúdo:** tendências regulatórias nacionais. Belo Horizonte: Instituto de Referência em Internet e Sociedade, 2021. Available at: https://bit.ly/3xjAUka. Accessed on: 6 Sep. 2024.

[2] SILVA, Fernanda dos Santos Rodrigues; GERTRUDES, Júlia Maria Caldeira; SILVA, Rafaela Ferreira Gonçalves da. **Regulação de plataformas e devido processo na moderação de conteúdo:** perspectivas em 5 continentes. Belo Horizonte: Instituto de Referência em Internet e Sociedade, 2024. Available at: https://irisbh.com.br/publicacoes/devido-processo-na-moderacao-de-conteudo-em-5-continentes/. Accessed on: 6 Sep. 2024.

[3] PEREIRA, Ana Bárbara Gomes. SILVA, Fernanda dos Santos Rodrigues; GERTRUDES, Júlia Maria Caldeira; SILVA, Rafaela Ferreira Gonçalves da. **Cartilha de Enfrentamento à Desinformação em Redes Sociais.** Belo Horizonte: Instituto de Referência em Internet e Sociedade, 2024. Available at:https://irisbh.com.br/wp-content/uploads/2024/08/Cartilha-de-Enfrentamento-a-Desinformacao-em-Redes-Sociais.pdf. Accessed on: 6 Sep. 2024.

content moderation can be managed more effectively and transparently in the Brazilian digital context. This research is a necessary step towards expanding freedom of expression online, user safety and reliability of digital platforms.

# Introduction

The lack of information about the history of content moderation decisions on digital platforms, especially social media, prevents an adequate view of the main problems faced by users during this procedure. Despite the existence of transparency reports, most of them only serve to count total numbers, such as moderated content, resources, and compliance with court decisions. The specific situations and decision-making rationales guiding the platform's actions are not disclosed.

This opacity, however, does not seem to help any of the parties involved. On one hand, users are unsure about how the moderation procedure will take place and how the platform acts in specific cases; on the other hand, platforms suffer from speculation around how they operate, which can lead to distrust among the public. Although fully disclosing all content moderation procedures and mechanisms could also have negative effects, allowing malicious individuals to exploit that information to circumvent the system, a minimum level of transparency and adherence to due process seem to be a way to balance the interests of both sides.[4]

As presented in a research IRIS published in 2024, different regulations aimed at platforms around the world have brought minimum parameters that fit the idea of due process.[5] Duty to notify moderation action, deadline for challenge and response by the company, among others, are some of the possibilities that international standards have foreseen to provide legal certainty to the scenario. However, fundamental information to help improve this activity across platforms and predict more objective policies remains hidden: what are the **main complaints from users regarding content moderation on social media? Our study points to scientifically based evidence on this question.**

[4] SILVA, Fernanda dos Santos Rodrigues; GERTRUDES, Júlia Maria Caldeira; DUTRA, Luiza Correa de Magalhães; SILVA, Rafaela Ferreira Gonçalves da. **Guia Informativo:** Devido Processo na regulação da moderação de conteúdo ao redor do mundo. Belo Horizonte: Instituto de Referência em Internet e Sociedade, 2023. Available at: https://bit.ly/3smC0i0. Accessed on: 6 Sep. 2024.

[5] SILVA, Fernanda dos Santos Rodrigues; GERTRUDES, Júlia Maria Caldeira; SILVA, Rafaela Ferreira Gonçalves da. **Regulação de plataformas e devido processo na moderação de conteúdo: perspectivas em 5 continentes.** Belo Horizonte: Instituto de Referência em Internet e Sociedade, 2024. Available in: https://irisbh.com.br/publicacoes/devido-processo-na-moderacao-deconteudo-em-5-continentes/. Accessed on: Aug 30, 2024.

## 1.1. How to identify the main complaints from users about the content moderation procedure on social media platforms?

As already mentioned, there is currently no official repository of decisions and/or complaints about the online content moderation procedure. The absolute numbers presented in transparency reports do not allow a more detailed visualization of the main problems faced by users, which could contribute to an improvement in the performance of the activity. With this in mind, it was necessary to find a public access platform that people would adopt as their main tool to express their criticism. In Brazil, this is often the platform ***Reclame Aqui*** ("Complain Here"), which, as the name suggests, allows consumers to make complaints about goods sold or services provided by certain companies. Its role is so relevant in the consumer area that, legally, courts across the country have recognized its validity as evidence in cases involving consumer rights.[6]

In the academic area, studies address different issues related to the use of this resource. With regard to consumers, there is research that explores how the platform can promote their empowerment, albeit with restrictions.[7] As for companies, research analyzed their *ethos* discursive in responding to complaints[8] and the role of the platform in reputational and image management.[9] And finally, regarding Reclame Aqui itself, it is possible to find works that analyze the evolution of the nature of the platform.[10] One way or another, the fact is that this space has been recognized as important for understanding the presentation of complaints by citizens regarding goods and services companies.

**With the choice of *Reclame Aqui* as the portal to verify users complaints regarding the**

[6] BRAZIL. Tribunal de Justiça de Minas Gerais (10ª Câmara Cível). Agravo de Instrumento nº 1.0000.22.069254-5/001. DIREITO CIVIL E PROCESSUAL CIVIL. AGRAVO DE INSTRUMENTO. AÇÃO DE OBRIGAÇÃO DE FAZER COM PEDIDO DE TUTELA PROVISÓRA DE URGÊNCIA. LIMINAR INDEFERIDA. RETIRADA DO NOME DA EMPRESA AUTORA DE RECLAMAÇÃO POSTADA NO SITE "RECLAME AQUI". REQUISITOS LEGAIS. NÃO CONFIGURAÇÃO. PONDERAÇÃO DE PRINCÍPIOS. PREPONDERÂNCIA DA LIBERDADE DE EXPRESSÃO E DE PENSAMENTO. Rapporteur: Des. Jaqueline Calábria Albuquerque, October 25, 2022. Available at: https://www5.tjmg.jus.br/jurisprudencia/pesquisaNumeroCNJEspelhoAcordao.do;jsessionid=B7BB6EDBF0ACA98E483A990DFA60BA76.juri_node2?numeroRegistro=1&totalLinhas=1&linhasPorPagina=10&numeroUnico=1.0000.22.069254-5%2F001&pesquisaNumeroCNJ=Pesquisar. Accessed on: 2 Sep. 2024.

[7] KOZINETS, Robert V.; FERREIRA, Daniela Abrantes; CHIMENTI, Paula. How Do Platforms Empower Consumers? Insights from the Affordances and Constraints of Reclame Aqui. **Journal of Consumer Research,** v. 48, n. 3, Oct./2021, pp. 428-455. Available at: https://academic.oup.com/jcr/article-abstract/48/3/428/6171149. Accessed on: 28 Aug. 2024.

[8] SÁ, Mirlene Batista. **O ethos discursivo em respostas de empresas no site Reclame Aqui.** Dissertation (Master of Arts) - Universidade Federal de Rondônia, Porto Velho, 2021. 97p.

[9] SEIVA, Carlos Eduardo Marques. **A apropriação da plataforma Reclame Aqui como balizadora na gestão de imagem e reputação organizacional na era da informação:** um estudo exploratório. Monograph (Administration Course) - Universidade do Vale do Rio dos Sinos, São Leopoldo, 2023. 75p.

[10] SELL, Cleiton Lixieski; CORRÊA, Jeano Saraiva. "Reclame Aqui" no Brasil: uma análise dos limites e possibilidades frente a perspectiva de um movimento social e seu viés em relação ao ativismo digital. **Revista Faculdade de Direito da UFC,** Fortaleza, vol. 39, nº 2, pp. 51-74, Jul./Dec. 2018. Available at: https://core.ac.uk/download/pdf/480546611.pdf. Accessed on: 28 Aug. 2024.

**content moderation procedure**, a new challenge emerged: the pages of digital social media platforms are not verified, meaning they do not have the Reclame Aqui verification seal, which, according to the site itself,[11] observes the following information, among others:

- "Having the company registered at the platform for free;
- Maintain a good reputation at *Reclame AQUI* (No Defined Reputation, Regular, Good, Great and RA1000);
- Have active service channels and CNPJ at *Reclame AQUI*;
- Having hired RA Brand Page, our brand positioning solution within the platform;
- Pass the monthly verification to earn the Verified Company Seal."

Even in the absence of the verification seal, it was possible to notice that **the social media pages to be analyzed were actually used by people who believed they were talking directly to the companies.** Although none of the chosen platforms had their page verified on *Reclame Aqui*, the page description, profile photo and/or the content and number of complaints indicated that it was a space related to the respective platform,[12] which instructed users to file their complaints right there. Thus, from a factual standpoint, **it was possible to conclude that the complaints found related to situations involving the mentioned social media platforms, especially from the users' perspective.**

Therefore, the methodology used to achieve the results of this research went through the following steps:

[11] RECLAME AQUI. Selo RA Verificada do Reclame AQUI verifica a existência e confiança das empresas. Available at: https://blog.reclameaqui.com.br/selo-ra-verificada-certifica-credibilidade-da-empresa/. Accessed on: 28 Aug. 2024.

[12] The platforms page on Reclame Aqui can be accessed at the following links: Instagram - https://www.reclameaqui.com.br/empresa/instagram/; Facebook - https://www.reclameaqui.com.br/empresa/facebook/; Twitter - https://www.reclameaqui.com.br/empresa/twitter/; TikTok - https://www.reclameaqui.com.br/empresa/tiktok/; and Youtube - https://www.reclameaqui.com.br/empresa/youtube_194558/sobre/.

### 1.1.1. How was data collected?

- Social media platforms investigated: Instagram, Facebook, Twitter, Tiktok e Youtube;
- Justification for platform selection: they are the most used by candidates for political office,[13] demonstrating their potential impact on one of the main processes of the democratic state of law;
- Expected collection sample: 250 complaints from each platform;
- Collection method:
  - The Reclame Aqui platform had some limitations, such as the fact that it did not allow selecting a time period to search for complaints and it was not possible to access beyond page 50 of complaints. As a result, and considering the large influx of new complaints every hour, we sought to collect complaints from the first 25 pages of each platform, containing 10 complaints each, which totaled 250 complaints per social network;
  - Complaints were selected from the "Content" category, which was generated by Reclame Aqui itself;
  - The collection was carried out by two researchers based on a printout of the complaint page, complaint link and content;
  - After collecting the selected complaints, we observed that some were repeated so the repetitions were excluded.
- Total sample analyzed: 1081 complaints
- Final sample per platform, excluding repetitions:
  - a) Instagram: 247; b) YouTube: 249; c) TikTok: 85; d) Facebook: 250; e e) Twitter: 250
- Limitations:
  - As it is the most recently created page on platform[14] and, consequently, with a lower number of complaints in the "Content" category, TikTok had a lower number

[13] MALI, Tiago. Sobe número de candidatos nas redes sociais; saiba as preferidas. **Poder 360**. 18 Aug. 2022. Available at: https://www.poder360.com.br/eleicoes/sobe-numero-de-candidatos-nas-redes-sociais-saiba-as-preferidas/. Accessed on: 6 Sep. 2024.

[14] The TikTok page on Reclame Aqui is listed as registered 4 years ago, while the others are between 7 and 13 years old.

of complaints analyzed;

- It is possible that complaints about the object of the research, namely, the content moderation procedures on the chosen platforms, were also classified under other categories besides "Content," and therefore were not collected in this study. However, this was a limitation imposed by the inability to review all the material on the site. Thus, in the tests conducted, the fact that the "Content" category returned main themes related to our objective demonstrated that it was a relevant and appropriate focus for this study.

### 1.1.2. How to analyze the data?

For data analysis, initial categories were created to classify complaints, based on a previous study [15] on the elements of a right to due process in content moderation. In other words, minimum parameters to ensure fairness and legitimacy in the platforms' moderation procedures toward users were the lens through which the complaints were analyzed. Considering that content moderation can involve both intervention in third-party content, such as removal of posts and suspension of accounts, as well as content recommendations in the feed, we chose to limit the scope of the research to analyze only the first case, namely, **complaints about post removal procedures and account suspension/blocking.**

Therefore, the first version of the list of categories was tested by three researchers from the team in two rounds, with meetings being held in each of them to discuss the classification and reflect on the need to change the initial list. At the end of this stage, the second version of the list of categories was as follows: *absence of appeal/challenge/review tools; inaccessible platform design; unaddressed challenge to a moderation decision; user was not informed about the moderation decision but noticed content removal; platform informed about the content moderation decision but did not provide justification; not about moderation; about moderation, but not related to our object; unable to determine if it is moderation; and others.*

In summary, the procedure for analyzing complaints in light of due process occurred as follows:[16]

[15] SILVA, Fernanda dos Santos Rodrigues; GERTRUDES, Júlia Maria Caldeira; SILVA, Rafaela Ferreira Gonçalves da. **Regulação de plataformas e devido processo na moderação de conteúdo: perspectivas em 5 continentes.** Belo Horizonte: Instituto de Referência em Internet e Sociedade, 2024. Available at: https://irisbh.com.br/publicacoes/devido-processo-na-moderacao-deconteudo-em-5-continentes/. Accessed on: Aug 30, 2024.

[16] The three researchers on the team were divided into analyzing the platforms, which was done in three rounds: (I) a round of initial analysis of each platform by its respective researcher; (II) a second round in which each researcher reviewed the classification of complaints on two or one platform that they had not worked on before; and (III) a third round of review in which each researcher took a new platform that they had not worked on and analyzed only cases of divergence between the first two to make a final decision. Afterwards, everyone collectively presented their third round opinions, in order to resolve any disagreements and verify any need for adjustment.

Division of complaints for each researcher;

1. Initial analysis according to this division;
2. Review of the analysis carried out by another researcher;
3. Analysis of divergences in a third round.

After analyzing all the collected complaints, it was identified that the large number involving hacked accounts required reclassification as a category in its own right, which prompted a review of the categorization of all platforms by two researchers. In this sense, considering the total sample of 1,081 complaints and the object of the research, complaints that did not pertain to moderation (28.03%), those related to hacking (23.03%), and those in which it was not possible to determine if it was a case of moderation (7.4%) were also discarded. The result is a final sample of 449 complaints about moderation, which were effectively analyzed.

Of these, 54.34% were related to our research object – specifically, moderation procedures in cases of post removal and account suspension/blocking – while 45.66% were about other aspects of moderation. Thus, the analyzed complaints were classified as follows:

# WHAT ARE THE MAIN COMPLAINTS IN CONTENT MODERATION ON SOCIAL MEDIA PLATFORMS?

1

### Inaccessible platform design

Cases where it was not possible to find how to contest/review/appeal the decision

2

### Lack of tools for appeal/contestation/review

Cases where the platform did not provide a way to appeal/review/contest the decision

3

### Unanswered moderation contestation

Cases where an appeal was submitted one or more times, and the platform never responded, or where there was an unjustified delay in addressing the review request.

4

### User was not informed about the moderation decision but noticed the content was removed

Cases where the user realized that content was removed because they could no longer find it, but the platform did not notify them at any point.

5

### The platform informed about content moderation but did not provide adequate justification

Cases where the user complained that they did not receive an adequate explanation for why their content was removed.

6

### Others

Situations that question the conflict resolution process but do not fit into any of the categories listed.

7

### It’s about moderation, but not related to our subject

Cases of third-party moderation for harmful content; monetization issues; third-party moderation for false identity; content recommendations; and complaints about moderation/decisions, but not regarding procedures.

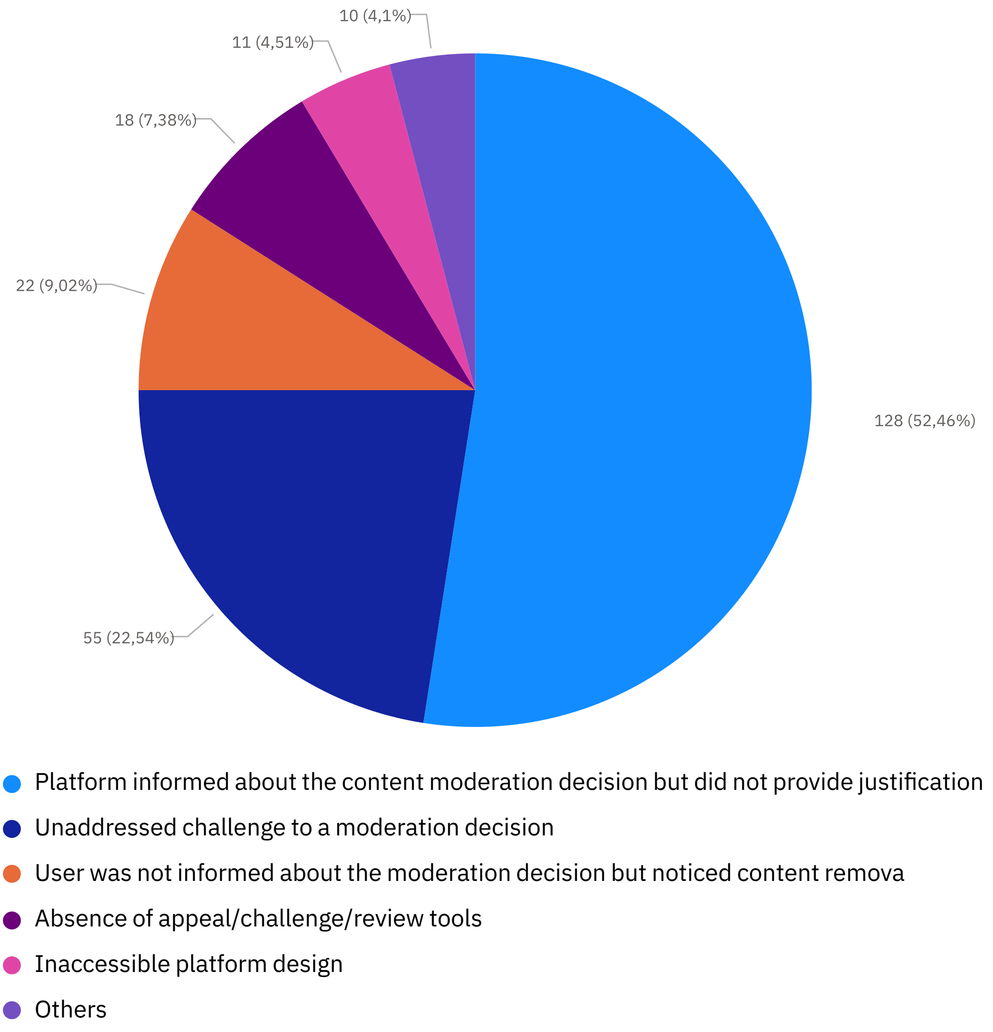
In relation to the 45.66% of the category concerning *moderation, but not directly to our research object*, it was decided to classify these cases separately at a later stage in order to better understand other content moderation issues that generated complaints. Thus, the subcategories were defined as follows:

- **Suspension due to linked account:** specific cases where accounts were moderated through blocking or suspension due to an alleged connection with another account on a different platform;
- **Generic complaint about platform moderation:** cases involving a) general complaints about how moderation functions (or fails to) on the platform, b) simple complaints about having been moderated, c) complaints about content recommendations on the platform (considering recommendations as a form of moderation), and d) complaints about moderation actions taken on third-party profiles;
- **Request for moderation of third-party content due to harmful content:** requests for moderation of profiles or specific posts based on harmful content;
- **Request for moderation of third-party content due to alleged false identity:** cases where profiles are accused of impersonating others;
- **Request for moderation of third-party content due to unauthorized use of images:** cases involving the unauthorized use of the complainant's or a third party's image;
- **Content recommendation/reach:** cases of dissatisfaction with the recommendation or reach of content produced by the complainant;
- **Restriction of platform functionalities:** cases where the complainant can no longer livestream, post stories, comment, or faces other functional restrictions as a result of a moderation decision by the platform;
- **Monetization:** cases involving monetization of profiles/accounts;
- **Account blocked due to platform age restrictions:** cases in which an account was moderated because of the age specified on the profile;
- **Request for moderation of ads:** cases involving the moderation of advertisements, distinct from content recommendations as ads are necessarily promoted or paid for;
- **Others:** includes residual complaints that do not fall into any specific subcategory.

The creation of these categories was part of a scientific effort to gain a deeper understanding of the primary types of user complaints regarding content moderation on social media. In some instances, clearly separating categories was challenging, as complaints often encompassed more than one issue. In these cases, the team focused on identifying the main concern of the complaint to classify it appropriately.

In order to avoid any type of ranking between the platforms analyzed, we chose to present the results jointly, without indicating the percentage of complaints for each platform. Section 2 presents the results of the sample complaints related to content moderation procedures, specifically regarding post removals and account suspensions/blocks. In section 3, you can follow the category analysis *it is about moderation, but not about our object.*

## 1.2. What are the main complaints about content moderation procedures regarding post removals and account suspensions/blocks?



Considering the categories related to our research object, namely **complaints about content moderation procedures concerning post removals and account suspensions/blocks,** it became clear that the majority of complaints *(52.46%)* were about *inadequate justification provided by the platforms for their moderation actions.* These cases include users who stated they did not understand the reasons behind the moderation decisions, as seen in the example below:

> *On [date hidden], my [platform name hidden] account was deactivated, and I received the following message we found severe or repeated violations of our community guidelines.*
>
> *However, I believe that all the videos posted on the channel and all my actions within the platform comply with the [platform name hidden] Terms of Service and Community Guidelines.*
>
> *Therefore,* ***I would like the platform to explain what these violations were and when they occurred,*** *and if possible, I request a review of the decision and the reactivation of my channel.*

In other words, many complaints involve concerns over the transparency and clarity of how content moderation decisions are justified — specifically, what content is allowed, the rules governing moderation procedures, and which specific terms of the community policies were violated. In the same category of inadequate justification, there were also cases where users complained about automated responses that failed to properly address the arguments presented in their appeal. The excerpt below illustrates this situation more clearly:

> *After the suspension, I actively sought to reactivate my account, but I have only received automatic and generic responses from the [platform name hidden], which do not offer any concrete clarification on how to resolve the situation. This lack of effective communication has left me completely helpless and without adequate means to contest the decision or resolve the problem satisfactorily.*
>
> *Therefore, I urgently request an impartial and transparent review of the situation, with the immediate reinstallation of my account on [platform name hidden] [insert username] or, if necessary, a detailed explanation of the reasons for the suspension and a clear path to resolve any possible wrongdoing.*
>
> *I hope for a quick and adequate response so that I can resume my normal activities without any further unnecessary losses.*

It is well known that digital platforms rely heavily on automated processes due to the sheer volume of content uploaded every minute.[17] However, complaints like these show that the lack of human oversight can potentially undermine users' rights to appeal. When a user submits an appeal, explaining why they believe they have not violated community policies, they expect their arguments to be considered in order to reverse the decision. Automated responses that cannot understand these arguments or provide a properly justified answer can lead to frustration and render the appeal process nearly ineffective. As a result, a crucial element of due process — the right to appeal — is compromised.

Although it is not possible to ignore the fact that there may be a lack of understanding by the user in relation to what is foreseen in the platform's community policies, which impacts their understanding of the basis for moderation decisions, this is a problem in which the platform also has responsibility. If the rules are hard to comprehend, it is the platform's role to adopt measures to simplify and clarify them, ensuring that users can understand and follow them more effectively.

The online content moderation process does not involve, in this sense, just the binary of removing or not removing certain online content, but the entire process of building the user-platform relationship and the rules and norms established to generate greater legal security, as well as a political-social construction of moderation actions - which involve subjectivities, discourses and freedoms. In other words, the way we deal with these issues affects communication and social interaction.[18]

The absence of essential elements in a well-founded moderation process, such as transparency, notification, and consistency, can lead to the perception of censorship. This concern becomes particularly relevant when *22.54% of the analyzed complaints involved unresolved challenges to moderation decisions.* This gap in the appeal process can fuel the perception that platforms act arbitrarily, highlighting the need for a more rigorous and transparent approach to content moderation.

In a context where many people rely on social media as their primary means of communication for work, the lack of a response to moderation appeals can leave them completely unsupported and potentially affect their earnings. The same applies to cases where *users discovered content had been removed but reported receiving no notification (9.02%).* Without communication from the platform, users are deprived of the opportunity to appeal through the proper channels, forcing them to seek alternative measures, such as using external platforms like Reclame Aqui. The following example illustrates this type of complaint:

[17] According to research carried out by Domo, a cloud computing service company, 66 thousand photos were shared on Instagram; 347.2 thousand tweets on the old Twitter – now X; and 1.7 million "pieces of content" on Facebook per minute in 2022. See more at: DOMO. Data never sleeps 10.0. Available at: https://web-assets.domo.com/miyagi/images/product/product-feature-22-data-never-sleeps-10.png. Accessed on: 17 June. 2024.

[18] GOLDMAN, Eric. Content moderation remedies. Michigan Technology Law Review, v. 28, p. 1-xx, 2021. Available em: https://repository.law.umich.edu/mtlr/vol28/iss1/2. Accessed on: 30 ago. 2024.

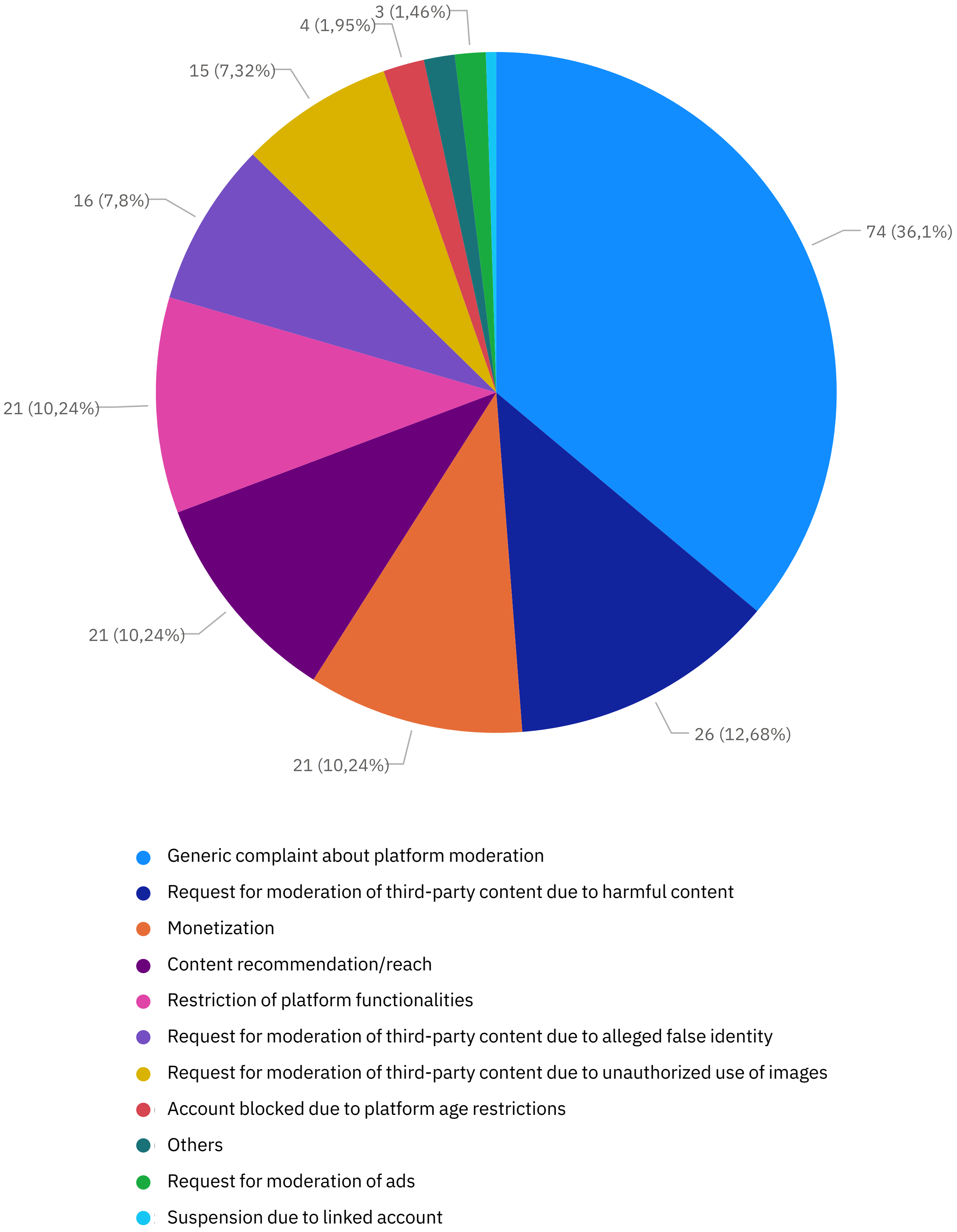
> *(...) Recently,* ***I discovered that several of my videos****, which accumulated more than 2 million likes,* ***were deleted without any prior notice or adequate explanation.*** *These videos represented hours of creative work, dedication and interaction with my followers, and their* ***abrupt removal*** *has negatively impacted my presence on the platform.*
>
> *Furthermore,* ***It is extremely disturbing to see that these deletions were carried out without my consent and, until now, I have not received any convincing justification from the support team of*** **[platform name hidden]*****. The lack of transparency and effective communication from the platform is unacceptable and disrespectful*** *to content creators who invest their time and effort in contributing to the [platform name hidden] community..(...)*

Indeed, content removal is already a sensitive measure that can generate frustration for users; however, carrying out such actions without any prior notification or justification tends to exacerbate the situation. As previously mentioned, in a reality where many people derive part or all of their income from the digital world, realizing that content has been removed without the platform's due notification breaks the trust established between the two parties. It creates an unmet expectation regarding the due process of informing users of any actions taken against them.

Lastly, although the percentages of complaints in the categories of *inaccessible platform design (4.51%)* and *lack of appeal/review mechanisms (7.38%)* are relatively low, they are still significant and point to important areas of concern. Continuous review of moderation processes, along with the optimization of platform interfaces and functionalities, is crucial not only for maintaining effectiveness but also for ensuring that platform actions are increasingly transparent and aligned with user needs.

It is essential that users have easy and intuitive access to effective mechanisms for requesting reviews and voicing disagreements. This approach ensures compliance with the principles of due process in moderation, allowing users to exercise their right to a broad defense and contest decisions fairly and transparently, without encountering unnecessary difficulties.

## 1.3. What are the other complaints about content moderation on social media platforms not related to post removals and account suspensions/blocks?



This section contains complaints identified within the category *It's about moderation, but not about our object*, which is equivalent to 45,66% of the total sample analyzed. These complaints involve various aspects of content moderation, including procedures, but **not specifically related to post removals or account suspensions/blocks.** As shown in the chart, the most common cases are *general complaints about the platform's moderation process (36.1%)* and a range of subcategories related to *requests for moderation of third-party content (30.24%).*

Regarding the first point, the general complaints, users expressed dissatisfaction with how moderation works (or doesn't work) on the platform, simple discontent with being moderated, complaints about content recommendations by the platform's algorithm, and complaints about moderation decisions applied to third-party profiles. Specifically, users' complaints about the lack of clarity in content recommendations – especially when they don't recall watching similar material before – highlight issues with the opacity of recommendation algorithms. While data collection and interaction analysis from partner companies are justified as ways to personalize user experiences, these complaints reflect a significant failure to deliver on the promise of effective and transparent personalization.

Similarly, complaints about moderation actions applied to third-party content (not the complainant's own) stood out. In this regard, it was noted that some complainants submitted complaints on behalf of channels or influencer accounts they followed that had been moderated, asking the platform to reverse the punishment. This indicates that platforms like Reclame Aqui are being used as a form of activism in support of content creators.

On the other hand, complainants also used this space to request moderation of third-party content. In this context, three main situations were identified:

12,68%

Requests for moderation of harmful content

Where people claimed to find content that violated the platform's community policies

10,24%

Accusations of false identity

Where individuals reported the creation of fake accounts impersonating them, and

7,32%

Complaints about unauthorized use of images

Where people indicated that third-party accounts were using their photos or those of their family or acquaintances without permission.

This use of external complaint platforms is notable, as all the platforms analyzed offer internal tools for reporting content that falls under these categories. In some cases, complainants mentioned they had already reported the issue through the platform but received no response. While it's possible that the content in question did not violate platform policies, the large volume of requests for moderation of third-party content highlights some form of inadequacy within the social networks — whether it's making their reporting mechanisms more accessible or reviewing internal policies to address gaps that allow harmful content, unauthorized image use, and the creation of fake accounts impersonating real individuals.

In third place, complaints related to monetization and content recommendation/reach each accounted for 10.24%. In the first case, these complaints referred to content moderation that limited access to monetization, reinforcing the role of platforms as sources of income for many individuals. Regarding content recommendation/reach, these complaints differ from general ones in that they specifically involve issues with the recommendation or reach of the complainant's own content. In some reports, users explicitly mentioned suspicion of being affected by *shadowbanning,*[19] where they noticed a reduction in the reach of their posts without receiving any notification from the platform about a decision:

> *Since October, as pointed out by other users, my name no longer appears in any search or hashtag.* ***Checking on websites that verify the presence of this type of censorship, I found that I am indeed "shadow banned" without having done anything abnormal, vulgar, or offensive.***
>
> *I thought this issue would resolve itself naturally over time, but to this day, I am still censored in searches, and* ***my content (artwork) simply doesn't reach users outside of my circle of followers,*** *which completely defeats the purpose of the social network.*
>
> *I tried contacting @Support (the support profile for [platform name hidden]) in English, and all I received were suggestions on how to behave correctly on social media, which is redundant since, as far as I know, I have not acted in any way that would warrant sanctions.* ***They were unable to inform me of the reason for the punishment.***
>
> *I request the removal of this negative measure against my account, information on how to lift it, or* ***at least some explanation as to why the account is under embargo.***

The difficulty in understanding what's happening (if anything is happening) and the reasons for such a decision are some of the related issues. Additionally, the inability to mount a defense or present counter arguments is a problem since users are not even informed that they have been moderated.

[19] See more about the concept in RADSCH, Courtney. Shadowban/Shadow Banning. In: BELLI, Luca; ZINGALES, Nicolo; CURZI, Yasmin (org.). **Glossary of platform:** law and policy terms. Rio de Janeiro: FGV Direito Rio, 2021.

Lastly, there are residual cases, including account moderation due to *age restrictions (1.95%), requests for ad moderation (1.46%), suspensions due to linked accounts (0.49%)*, and *others (1.46%)*. These highlight residual issues with social networks that could benefit from more effective communication methods and appeals processes on the investigated digital platforms.

iris

INSTITUTE
FOR RESEARCH
ON INTERNET
AND SOCIETY