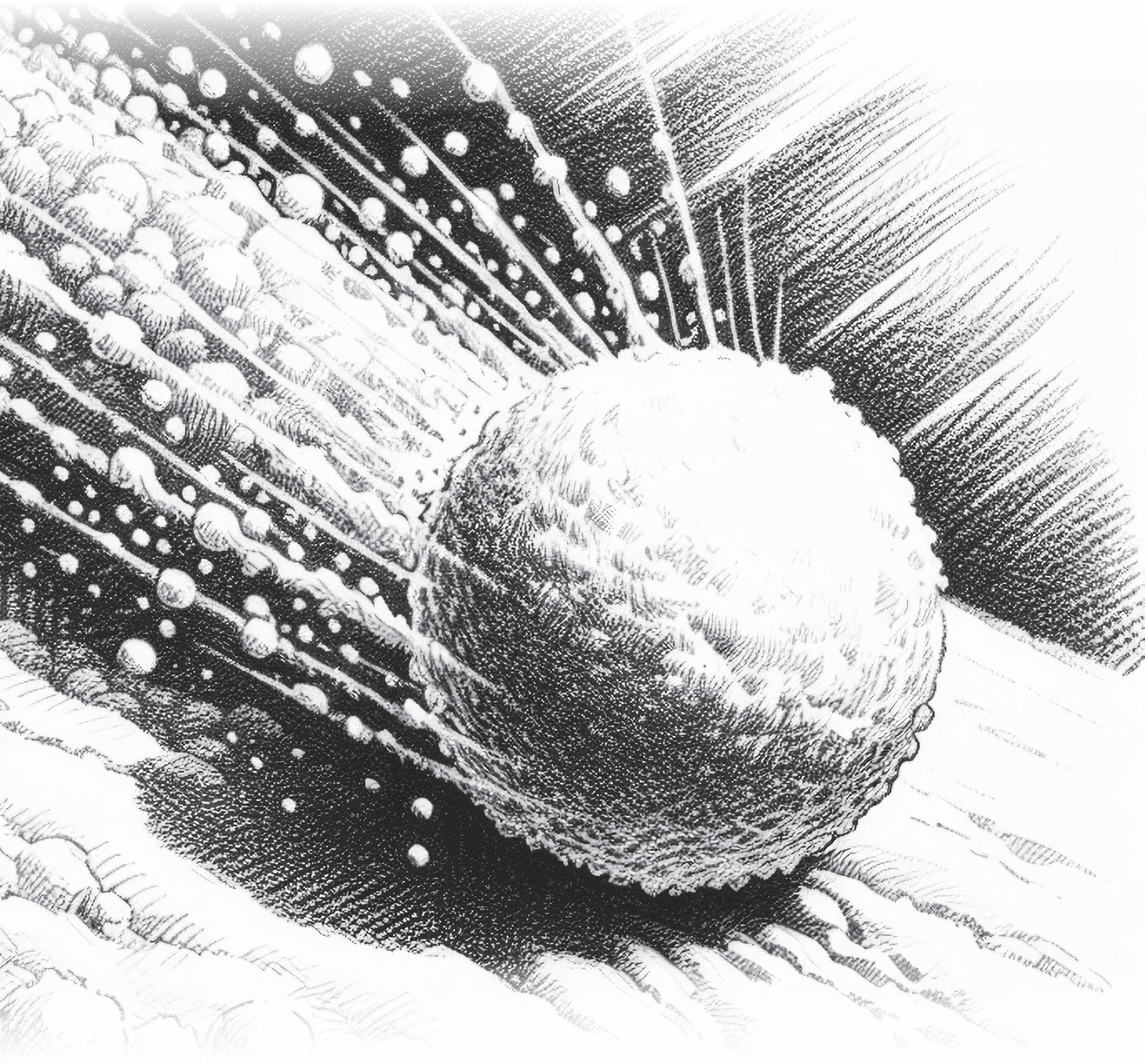
ANNUAL REPORT

MONITORING HATE SPEECH ON SOCIAL MEDIA

Racist, xenophobic, Islamophobic, antisemitic and anti-Roma contents

2024



GOBIERNO DE ESPAÑA
MINISTERIO DE INCLUSIÓN, SEGURIDAD SOCIAL Y MIGRACIONES
SECRETARÍA DE ESTADO DE MIGRACIONES

oberaxe
Observatorio Español del Racismo y la Xenofobia

Co-funded by the European Union

Catalogue of publications of the General State Administration
https://cpage.mpr.gob.es

Madrid, 2025

Author(s): Spanish Observatory on Racism and Xenophobia (OBERAXE)

Published by: Spanish Observatory on Racism and Xenophobia
Distribution: Spanish Observatory on Racism and Xenophobia
Calle Agustín de Betancourt, 11, séptima planta. 28003 Madrid

E-mail: oberaxe@inclusion.gob.es
Web: https://www.inclusion.gob.es/oberaxe/es/index.htm

Publication Identification Number (NIPO): 121-25-080-4

Design: DISEÑO GRÁFICO GALLEGO Y ASOCIADOS, S. L.
Layout: CYAN, Proyectos Editoriales, S.A.

Recommended citation: Spanish Observatory on Racism and Xenophobia [OBERAXE] (2025). *Annual Report: Monitoring Hate Speech on Social Media 2024*. Retrieved from: https://www.inclusion.gob.es/oberaxe/es/publicaciones/documentos/documento_0164.htm

CONTENT

List of tables

List of graphs

1. INTRODUCTION

The rise of social media has profoundly transformed the way people communicate, access information and engage in the public sphere. This transformation has also brought with it new challenges in terms of coexistence, respect and protection of fundamental rights. In particular, hate speech in digital environments represents a growing phenomenon that has a negative impact on social cohesion, the safety of vulnerable groups and society at large and the quality of public debate.

Given this reality, the need for institutional tools to monitor, analyse and provide information to combat hate speech on social media has become increasingly evident.

In response to this challenge, the Spanish Observatory on Racism and Xenophobia (OBERAXE), part of the State Secretariat for Migration of the Ministry of Inclusion, Social Security and Migration, implemented in 2020 a system for monitoring hate speech on social media works that allows for the analysis of the characteristics of this type of discourse, and its results are presented periodically in bulletins published on the OBERAXE website, initially on a bimonthly basis and, since September 2024, on a monthly basis.

In addition, the first annual report on the results of hate speech monitoring was published in 2023, which was a key step in consolidating a systematic and evidence-based approach to tackling hate speech. The report presented the conceptual, methodological and institutional basis for rigorous and sustained monitoring of hate speech, providing for the first time a detailed annual analysis of its scope, main themes and recurrent patterns.

The present report, corresponding to the year 2024, continues this line of work, maintaining the general approach of the previous year, while introducing methodological improvements and technical adjustments to refine the analysis and improve the detection of emerging phenomena.

One of the most important milestones for OBERAXE in 2024 was the signing, on 24 October, of the agreement between the Ministry of Inclusion, Social Security and Migration and LALIGA, which enabled the development of the FARO System.

The FARO System is the new methodology to be used by OBERAXE, from March 2025, for the real-time identification and analysis of hate speech content with racist,

xenophobic, Islamophobic, anti-Semitic and anti-Roma motivation, and which allows the incorporation of advanced artificial intelligence for the monitoring of social media, significantly increasing the number of potentially criminal hate speech content identified.

The FARO System incorporates the use of artificial intelligence technology, trained in LALIGA's Monitor for the Observation of Hate in Sport (MOOD), to the knowledge and experience that OBERAXE has in monitoring hate speech on social media.

The FARO System is the conjunction of the use of two tools (FARO Monitor and ALERTODIO) combined with a new working methodology that concerns both the identification of content and the analysis and presentation of results through a real-time data visualisation monitor. The FARO System data visualisation monitor is available for public consultation on the new OBERAXE web portal.

OBERAXE's systematic monitoring of hate speech on social media is carried out in close collaboration with the State Attorney General's Office and the State Security Forces and Corps. This cooperation is channelled through the Guardia Civil's Crime Response Teams (REDO) and the National Police's specialised Violent Extremism and Hate (EVO) teams, focusing on the detection and investigation of complex hate incidents and crimes.

In addition, OBERAXE's activity is supported by projects co-funded by the European Commission, such as the CISDO project, carried out in collaboration with the National Office for the Fight against Hate Crimes (ONDOD) of the Ministry of the Interior. This project aims to improve police capacities nationally and locally to prevent, identify and respond to xenophobic and racist incidents and hate crimes, and to provide better assistance to victims; the SCORE project, which aimed to create a coalition of European cities and local authorities for the promotion of inclusive sport, as well as the prevention and fight against racism, xenophobia and related intolerance in sport, and the REAL UP project, aimed at improving the capacities of state authorities to identify, analyse, monitor and evaluate online hate speech in order to develop and strengthen counter-narrative (upstander) strategies against hate speech motivated by racism, xenophobia, Islamophobia, anti-Semitism and anti-Roma sentiment.

On the other hand, the recent full entry into force of Regulation (EU) 2022/2065, known as the Digital Services Act (DSA), the DSA, as of 17 February 2024, represents a significant step towards regulating the liability of digital service providers, including digital platforms, search engines and other online providers. This regulation sets out clear obligations for these platforms to ensure a safe, transparent and competitive digital space, protecting users' fundamental rights and fostering innovation in the European single market.

At the European level, too, the renewed version of the Code of Conduct on Combating Illegal Hate Speech on the Internet+ was launched on 20 January 2025, signed with the European Commission and by major digital platforms such as Meta, X, YouTube, TikTok, LinkedIn and Microsoft, among others. This Code reinforces the voluntary commitments made in 2016 and complements the legal framework established by the Digital Services Act (DSA), promoting clear policies prohibiting hate speech, effective reporting mechanisms for European users, diligent review of reports and transparency in human and automated moderation processes. It also establishes a system of cooperation with specialised civil society organisations, enabling knowledge sharing and improving the identification and analysis of hate speech online.

In Spain, the Comisión Nacional de los Mercados y la Competencia (CNMC) was designated national coordinator for the implementation of the Digital Services Act on 24 January 2024, assuming a key role as the single point of contact and responsible for overseeing the correct implementation of the DSA at national level.

Furthermore, the CNMC remains committed, actively participating in the working groups of the European Digital Services Board and promoting dialogue with national institutions, including its collaboration with OBERAXE and its participation in the "Agreement to Cooperate Institutionally against Racism, Xenophobia, LGTBIphobia and other forms of intolerance". OBERAXE aspires to be nominated as a "trusted flagger" by the CNMC in 2025, given its work and experience in detecting and combating hate speech and its involvement at international and European level, among others, in the European Commission's High Level Group on Hate Crime and Hate Speech.

OBERAXE welcomes the recent appointment of the Prosecutor María Teresa Verdugo Moreno as Independent Authority for Equal Treatment and Non-Discrimination, in compliance with Law 15/2022 on equal treatment and non-discrimination, as it will facilitate the fight against many of those discriminatory contents that do not qualify for criminal prosecution. Progress can also be made in the streamlined processing of administrative sanctions for perpetrators, which will discourage this type of behaviour. Likewise, the actions of the Consejo Superior de Deportes, by virtue of Law 19/2007 against Violence, Racism, Xenophobia and Intolerance in Sport, which has had some exemplary judgements.

Spain has proven to be a benchmark in the fight against hate speech through the work of various institutions, including OBERAXE. In fact, the continuous monitoring and collaboration with public and private actors has been recognised, among others by the Deputy Secretary General of the Council of Europe, Bjørn Berge, who highlighted

in December 2024 the relevance and impact of the work carried out by OBERAXE in this field.

In addition, in October 2024, a conference on counter-narratives and alternative narratives was organised by OBERAXE in collaboration with the Council of Europe, attended by some twenty civil society organisations and other institutional actors working for the inclusion of immigrants and against discrimination.

The main goal of the workshops was to provide the participating organisations with knowledge and tools for the formulation of alternative counter-narratives and narratives to combat racist and xenophobic hate speech and to raise awareness through communication campaigns, with a human rights-based approach. They addressed concepts such as the impact of hate speech and the analysis of discriminatory narratives in the context of Spain

In conclusion, the situation of hate speech on social media is complex and requires effort and resources, as well as approaches from different areas to combat it: criminal and administrative legislation, counter-narrative strategies, awareness-raising and training, monitoring and analysis of the situation.

The integration of technological resources such as artificial intelligence, the strengthening of national and European institutions, and cooperation with platforms and civil society are the pillars on which an effective response to this challenge must be based. Hate speech monitoring not only allows us to diagnose the situation, but also to contribute with information for the design of policies and strategies that contribute to ensuring a digital environment free of discrimination, intolerance and hostility towards people of foreign origin, thus favouring their inclusion in society and social cohesion.

The overall goal of the OBERAXE monitoring system is to identify, characterise and evaluate the presence of hate speech on social media, with the aim of:

- Contributing to empirical knowledge on its magnitude and evolution.
- Detecting emerging trends and thematic hotspots of hostility.
- Informing and guiding public policy on prevention, awareness-raising and intervention.
- Strengthening institutional collaboration and collaboration with digital, academic and civil society actors.

The system encompasses both explicit hate speech and more subtle, implicit or codified forms, recognising that discriminatory discourses take multiple formats, degrees of intensity and levels of visibility.

OBERAXE's monitoring focuses on the search, collection, analysis and notification to digital platforms of content that constitutes hate speech with racist, xenophobic, Islamophobic, anti-Semitic and anti-Roma motivations. Such content may constitute a criminal offence, infringe administrative regulations or violate the rules of use of the platforms themselves.

The scope of action covers only speech directed at individuals or groups on the basis of their ethnic, national or religious origin. Due to their specific vulnerabilities, particular attention is paid to vulnerable groups such as immigrants, unaccompanied children and youth and refugees.

Specific Goals

The specific goals of monitoring are twofold:

1. **Evaluating the response of platforms:** This involves analysing how digital platforms manage the removal of reported illegal hate speech content, in line with the commitments made under the EU Code of Conduct and the obligations established by the Digital Services Act (DSA).

Platform moderation is based on two fundamental pillars:

- The removal of illegal content according to the national legislation of the EU Member States, as provided for in the DSA.
- The removal of content that violates each platform's internal rules of use, a voluntary action in response to their own commitments, including adherence to the European Code of Conduct.

2. **Analysing hate speech and detecting trends:** Beyond reporting and evaluation responses, a detailed analysis is conducted on racist, xenophobic, Islamophobic, anti-Semitic and anti-Roma hate speech in Spain, thus contributing to a better understanding of the situation and facilitating the orientation of the design and implementation of public policies.

It should be noted that OBERAXE maintains active collaboration with digital platforms, establishing continuous communication channels that include regular meetings for the exchange of information, analysis and discussion on the identification and moderation of hate content.

3. METHODOLOGY

This report presents the results of OBERAXE's social media hate speech monitoring exercise for the year 2024. It continues the work initiated in previous years, especially in the 2023 Monitoring Report, which presented the conceptual foundations, objectives and methodological structure of the system.

The monitoring system developed by OBERAXE is based on a mixed methodology, allowing for a multidimensional approach to the complexity of the phenomenon, incorporating quantitative, linguistic, contextual and socio-cultural elements.

The methodology applied is based entirely on a manual and systematic search for content, carried out on a daily basis by a team of monitors. This team carries out direct observation of open profiles on social media, as well as the monitoring of debates, current news, viral publications and the use of a regularly updated glossary of terms and expressions. This strategy makes it possible to capture not only overtly hostile messages, but also more subtle or coded forms of hate speech.

The complexity of discriminatory discourse and its dependence on the socio-political context make manual review essential to ensure rigorous analysis, capable of identifying nuances, cultural references, euphemisms or intersectional elements that might go unnoticed with automated methodologies.

All identified contents are systematically registered in the ALERTODIO application, developed in collaboration with the Polytechnic University of Valencia. This tool facilitates a homogeneous, detailed and structured recording of information, allowing the annotation of key variables [1] such as the motivation of the discourse, the context of publication, the type of content, the language used and the target group addressed.

Once content that could constitute illegal hate speech or violate EU rules on digital platforms has been registered, it is reported to the platforms in a staggered

1. Definition of variables: see chapter 3.2.1. of the Annual Social Media Hate Speech Monitoring Report 2023

procedure. Then, from the initial notification made as a normal user, the response of the platforms is systematically monitored (whether or not they remove the reported content), with reviews after 24 hours, 48 hours and one week. If the platform has not removed the content after this period, the content is reported again, this time as a *trusted flagger*[2]. This process makes it possible to assess the response of the platforms and evaluate their commitment to moderating discriminatory content.

For a more detailed explanation of the system of identification, classification and notification of content, as well as the criteria used, the reader is referred to the Annual Report on Hate Speech Monitoring on Social media 2023 where the OBERAXE methodology for monitoring hate speech is described in depth, and which is maintained in this report.

2. Definition of trusted flagger: Trusted flaggers are individuals or entities that have been accredited by the data hosting service provider as having the necessary qualifications or expertise to report hate speech content. This accreditation is granted to those who are active in anti-discrimination issues and have the necessary experience in this field.

4. RESULTS

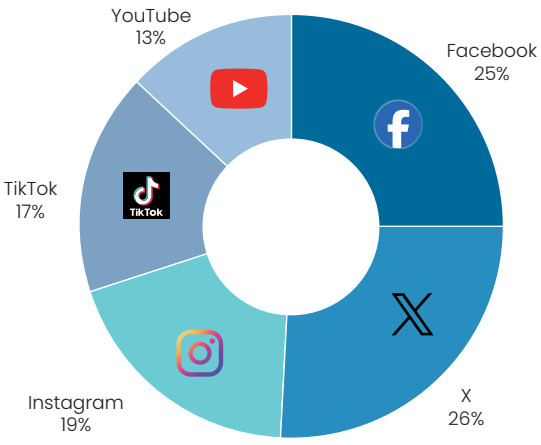
4.1. Monitored and Reported Content and Reaction of Social media

In 2024, 2,870 pieces of content were identified as racist, xenophobic, anti-Semitic, anti-Roma or Islamophobic hate speech. Such content could constitute a criminal offence, an administrative offence or violate the rules of conduct of internet platforms. These instances were reported to the five social media that were monitored (Facebook, X, Instagram, TikTok and YouTube).

The distribution of communications made to each platform (Graph 1) reveals a clear predominance of those made to X, with 758 cases (26% of the total). This is followed by Facebook with 727 cases (25%), Instagram with 538 (19%), TikTok with 478 (17%) and YouTube with 369 (13%). The discrepancy in the volume of content reported is primarily attributable to the varying degrees of difficulty in identifying content on each social network.

A total of 1,010 pieces of content were removed by platforms, representing 35% of those notified to them. Of all reported content, only 9% (272) was removed when reported through a normal user profile, while 26% (738) was removed after being reported through a trusted flagger. These data show a greater effectiveness in the removal of content when the notification is made through official channels or recognised as a reliable reporter.

Graph 1. The proportion of communications directed to each platform



Graph 2. **The percentage of content removed by the passage of time since notification to all monitored platforms in the year 2024 is presented herewith**

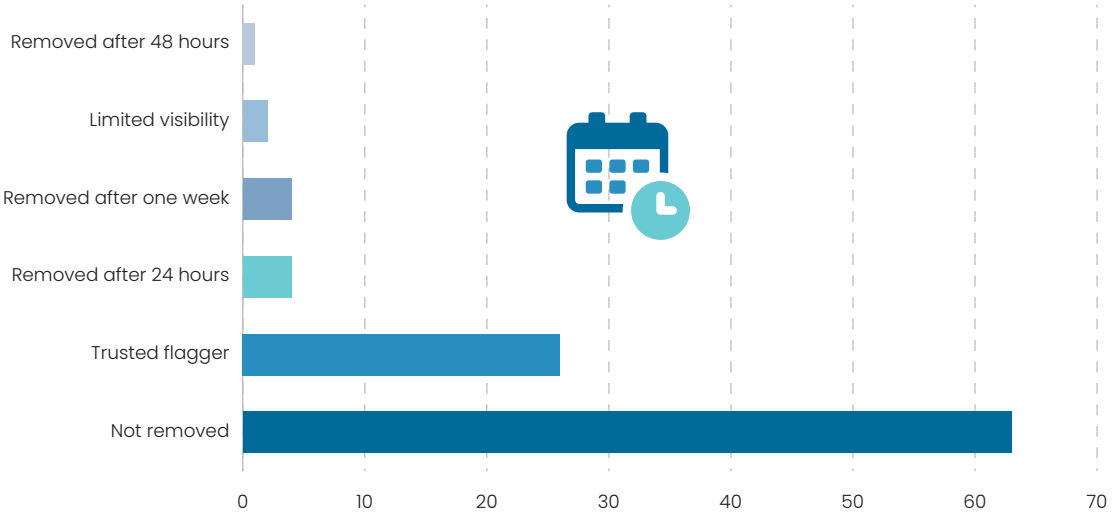


Table 1. **The percentage of content removed according to the time elapsed since the notification was issued and by platform, 2024**

	Total, content removed %	Removed after 24 hours %	Removed after 28 hours %	Removed after one week %	Removed *Trusted Flagger* %	Not removed %
Facebook	29%	1%	1%	5%	23%	71%
X/Twitter	14%	4%	1%	3%	7%	86%
Instagram	49%	2%	2%	7%	37%	51%
TikTok	69%	15%	3%	5%	47%	31%
YouTube	27%	0%	0%	1%	25%	73%
Total	35%	4%	1%	4%	26%	65%

However, the removal rate for all platforms is very low and has decreased by 14 percentage points compared to 2023.

The most efficient platform in terms of content removal is TikTok, which removed 69% of the total content reported to it. This is followed by Instagram (49%), Facebook (29%), YouTube (27%) and X (15%).

4.1.1. Characteristics of Content Removed at 24 Hours, 48 Hours and One Week

The platforms' response time to reported content is shown in Graph 2, which reveals that most removals occurred when content was reported as a trusted flagger (26%). Table 1 illustrates the efficiency and speed of content removal at 24h, 48h, per week, or via trusted flagger. It can be observed that the results differ between the five platforms. However, it can be concluded that all of them are more responsive to content removal when the trusted flagger route is used.

The data shows that the platforms' response to notifications made from normal user profiles is not very effective, especially in the first 24 and 48 hours, relevant periods to minimise the impact of hate speech. The low rate of immediate removal, only 4% within 24 hours, reveals weaknesses in moderation systems. This poor initial reaction from platforms makes it easier for content that dehumanises, promotes stigma or incites violence to remain visible and circulate widely, affecting target groups in particular. These dynamics can contribute to the normalisation of hate speech online, underlining the need to strengthen platforms' moderation and response mechanisms.

Of the content reported by regular users, TikTok is the social network that removed the most content within the first 24 hours (15%), followed by X (4%), Instagram (2%), Facebook (1%), and lastly YouTube, which has only removed 0.3% of its content within 24 hours.

36% of the content removed within the first 24 hours was found to contain language that was dehumanising, demeaning or aggressive in nature. The primary demographic targeted by these content removals within 24 hours were individuals from North Africa (43% of cases), with public safety representing the prototypical episode (40%).

The group of platforms has removed 1% of the content within 48 hours of being reported. In the content removed within this period, 53% of the content that dehumanises or seriously degrades, and 50% of the content that promotes discredit based on personal attributes, predominated. The principal prototype episode persists in the form of public safety, which is indicated in 32% of the removed notifications. Furthermore, the target group is also comprised of individuals from North Africa, representing 65% of the total.

The platform that removed the most content within 48 hours was TikTok, which deleted 3% of the reported content, followed by Instagram (2%), X (1%), Facebook (1%) and YouTube (0.3%).

Regarding content removed within a week, the group of platforms has taken down 4% of the reported content, with Instagram being the platform that removed the most during this time frame (7%), followed by Facebook (5%), TikTok (5%), X (3%) and YouTube (1%). It is noteworthy that 28% of the cases removed within a week are not linked to any prototypical episode, and that 59% of the communications removed within this timeframe contain explicit aggressive language.

4.1.2 Content Removed as Trusted Flagger

The "trusted flagger" route continues to establish itself as the most effective mechanism for the removal of hate speech content by platforms. Of the 2,870 notifications, 26% were deleted after being communicated through this channel, in contrast to the 9% effectiveness observed when the notification was made from a normal user profile. The difference in removal rates by route is particularly significant, showing that the platforms give priority to trusted flaggers.

When the data is disaggregated by platform, significant differences in the level of effectiveness of content removal can be seen. TikTok is the most efficient platform in terms of trusted flaggers, with 47% of content removed via this route. It is followed by Instagram with 37%, YouTube with 25%, Facebook with 23% and, lastly, X, which has a 7% removal rate through this channel.

4.1.3. Characteristics of Non-Removed Content

The percentage of content that was not removed was 65% (1860 cases). This percentage comprises 2% of notifications for which visibility has been reduced by the social network X. This mechanism was established by the end of 2023 as a positive action to diminish the effect of hate content, which, although it continues to circulate on the network, is less visible to users.

Despite the rules and mechanisms established by the platforms in the framework of the Code of Conduct and the regulations established by the DSA, the removal of hate speech content is still insufficient considering that 96% of the communications violate the very rules established by each of the platforms. The qualitative analysis of the 1,860 items of content not removed shows the following:

- In 38% of cases, discrediting of personal attributes is promoted.
- 36% dehumanise or severely degrade the target group.
- 28% incite violence by direct or indirect threats.
- 17% call for the expulsion of persons of foreign origin.
- 633 cases targeting North Africans have not been removed.
- The narrative of linking public safety to target groups is predominant.

4.2. Characteristics of Hate Speech on Social Media

4.2.1. Target Groups

One of the fundamental axes of OBERAXE's monitoring of hate speech on social media is to analyse who the speech is directed at. This identification makes it possible not only to map patterns of hostility towards people of foreign origin, but also to guide actions to prevent the dynamics of discrimination.

In 2024, data collected by OBERAXE reveals that in almost eight out of ten pieces of content reported to platforms, the message was addressed to a specific target

group (80%), compared to 20% of messages addressed to specific individuals. This trend confirms the structural nature of hate speech, which tends to reinforce stereotypes and fuel hostile attitudes towards different target groups.

When analysing the target groups of hate speech in detail, hostility towards North Africans stands out, accounting for 35% of all identified hate speech. This is followed by Africans and people of African descent at 24% and the generic category of migrants at 21%. This last figure is particularly significant, as it is evidence of a trend towards widespread discrimination on the grounds of other origin or nationality.

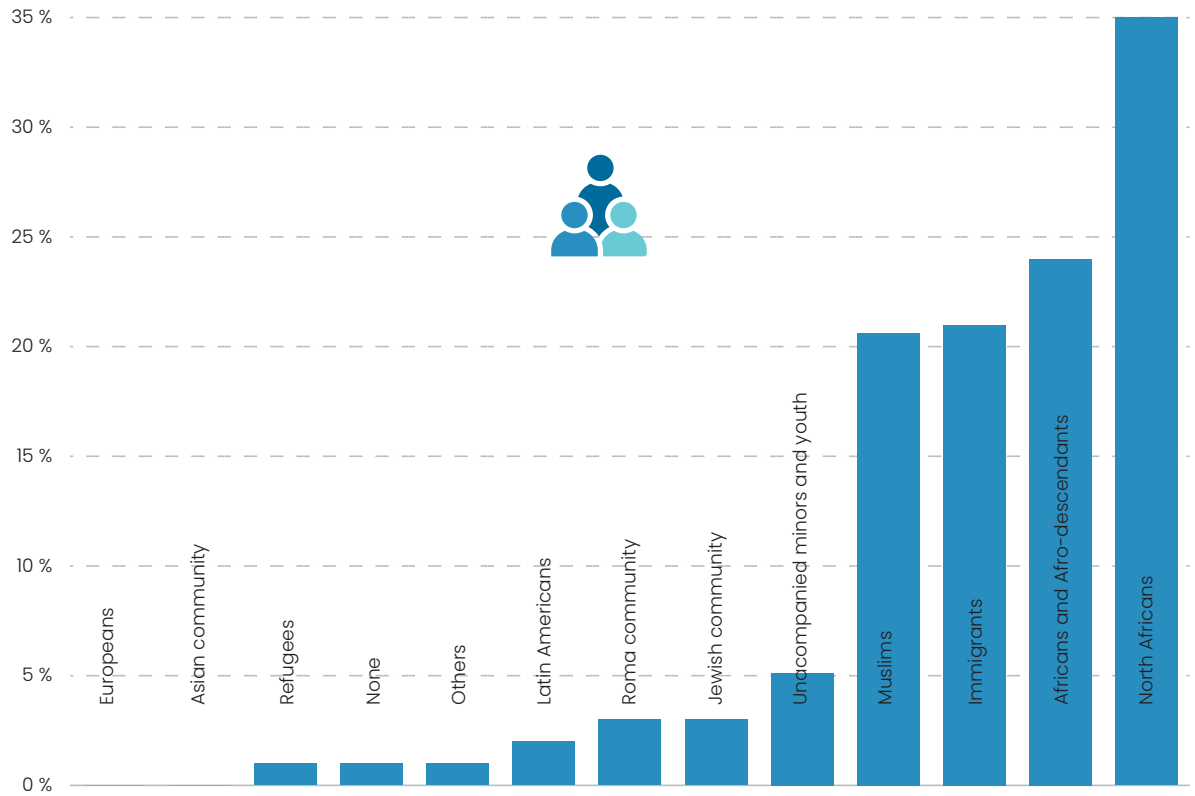
On the other hand, Islamophobic discourse is also significant. Muslims are the fourth most affected group, accounting for 21% of the total. These figures reflect a continuity in the dynamics of stigmatisation towards communities with visible religious markers, which are often instrumentalised through narratives of insecurity, criminalisation and other forms of symbolic exclusion.

Another relevant aspect of the analysis is the impact of the geopolitical context on the evolution of hate speech. The conflict in the Middle East (Israel and Palestine) has had a direct effect on the volume of anti-Semitic messages on social media. Hate speech towards the Jewish community stands at 3%. Similarly, an increase in hostile messages towards Muslims has been detected in connection with this conflict.

Hate speech directed towards unaccompanied children and adolescents is also noteworthy, accounting for 5% of the notifications. This data reflects how unaccompanied minors, especially those in situations of greater vulnerability, become the object of stigmatisation and discrimination, demonstrating the persistence of prejudices that dehumanise them and expose them to situations of risk, hindering their protection and well-being.

Other groups affected, although to a lesser extent, include the Roma community (3%), Latin Americans (2%), refugees (0.52%), Asians (0.42%) and Europeans (0.31%).

Graph 3. **The proportion of hate speech directed at each target group is presented herewith**



Looking at the yearly evolution of the target groups (Graph 4) during the first months of 2024, hate speech showed significant fluctuations. Particularly noteworthy is the high level of hostility towards the Jewish community, which peaked in January at 28%, with secondary peaks in July (13%) and October (13%). Content towards Latin Americans and Africans and Afro-descendants was also persistent, reaching high values of 20% and 5%, respectively, in December.

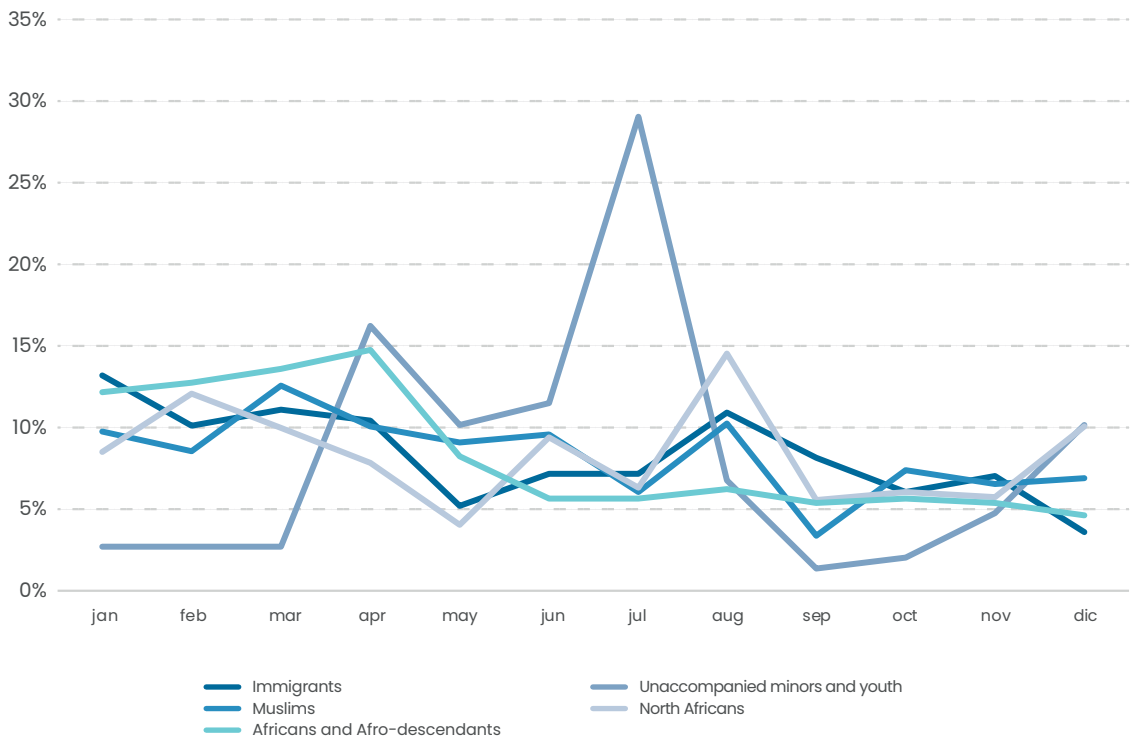
On the other hand, hostility towards unaccompanied children and young people increased sharply in April (16%) and July (29%), reflecting a worrying focus on this vulnerable group in those months. This pattern may be linked to specific media coverage or institutional measures for managing child migration, such as the distribution of unaccompanied minors among the Autonomous Communities, which provoked negative reactions on social media. Both months also saw spikes in discourse against other migrant groups, reinforcing the hypothesis of an intensified hostile narrative around migration issues driven by political agendas.

High levels of rejection towards the Roma community were also observed with significant peaks in April (24%) and June (17%), as well as towards North Africans, with peaks in August (15%) and December (10%).

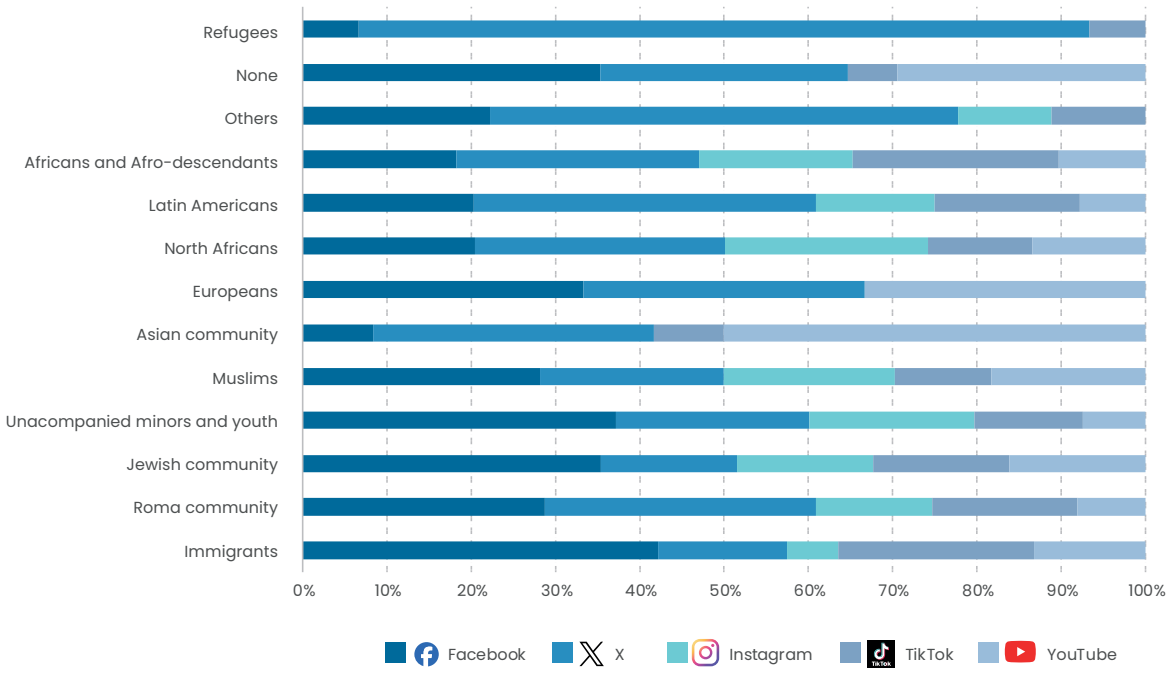
In relation to hostility directed at immigrants in general, although the percentages decreased progressively from January (13%) to December (4%), this group remains one of the most affected consistently throughout the year. Similarly, Islamophobia and hostility towards Muslims showed fluctuations, peaking in March (13%) and August (10%), coinciding with the celebration of Ramadan and specific events detailed in the section on prototypical episodes.

With regard to the target group at which the hate speech is directed, depending on the platform where it is predominantly disseminated, it is observed that this varies, although no causal relationship has been identified. This may be due to the audiences and dynamics specific to each platform (see graph 5). On the platform X, hate speech mainly targets North Africans (30%), followed by the Roma community (32%); Africans and people of African descent (29%). On Facebook, hate speech is

Graph 4. **Evolution of the main target groups in 2024**



Graph 5. **Prevalence of hate speech target groups on each social network**



most often directed at migrants (42%), unaccompanied children and adolescents (37%) and the Jewish community (35%). We should also note that on Instagram we find a high percentage of hate speech targeting North Africans (24%).

4.2.2. Types of Hate Speech

Among the types of discourse observed (Table 2), dehumanisation or severe degradation is present in 37% of the reported communications. This contributes to the violation of human dignity. Some examples of reported cases include the following: *"This remedy works like a charm for those ANIMALS. (GIF of a shotgun being loaded)"*; or *"The only way is for all of us to come together and start a hunt against these rats..."*.

Likewise, discredit based solely on personal characteristics of the group, or without providing any argument beyond belonging to it, appears in 32% of the hate content analysed. Examples of such content include: *"Immigrants and monkeys back to Africa, their habitat"* or *"They get that colour from the mud"* reduce people to biological stereotypes, ignoring their identity and humanity. This type of discourse not only reflects racial and xenophobic prejudice, but also fosters intolerance and exclusion by promoting the idea that certain physical characteristics justify discrimination. In addition, social stigmas are fostered which impede social cohesion.

Similarly, results show that in 29% of the monitored incidents, there is incitement to violence, with direct or indirect threats against migrants and/or people of foreign origin ("*Another one who doesn't eat ham — they should hang him in the town square as a warning,*" *"I've made up my mind. We have to get rid of them.*") Furthermore, in 22% of the cases, the group targeted by the hate speech is portrayed as a threat due to their actions ("*These are the ones the 💩💩💩💩💩 government funds and brings over in boats and planes to destroy the country.*" *"Watch out for the migrant kids and panchitos. 👀*")

Furthermore, 15% of the cases call for groups to be deported ("*They should be kicked out*"; *"Migrant kids out!!! 🐼🐼*"), which may lead to an increase in violent acts.

Also 5% of the content analysed praises those who attack the target group, thus legitimising violence and discrimination. Such messages reinforce intolerance and convey the idea that attacking target groups is acceptable. As a consequence, social polarisation is aggravated and coexistence is put at risk.

Regarding the distribution of target groups according to the type of hate speech, both Graph 6 and Table 3 reveal that content involving dehumanization prevails, particularly toward Asian individuals (58%), African and Afro-descendant individuals (48%), and Semitic and Jewish individuals (48%). Conversely, in discourse directed towards the Roma community, discrediting is a prevalent strategy employed in 63% of instances, based

Table 2. **Distribution of reported hate speech types**

Types of discourse	(n)	(%)
Inciting violence by direct or indirect threats	843	29
Dehumanises or seriously degrades	1058	37
Praises those who attack the target group	143	5
Calls for groups to be deported	430	15
Promotes hate on the basis of personal attributes	911	32
Presents the group as a threat due to its actions	642	22

on the ascription of personal or collective attributes, and which reflects the social stigmatisation experienced by this group.

It is noteworthy that 32% of the content directed at the Jewish community offers praise to the user who makes the comment, given that the majority of it alludes to Nazism. With regard to children and unaccompanied youths, 32% of the notifications presented these groups as threats to society.

Conversely, an examination of the content of hate speech was conducted to ascertain whether other vulnerable groups were referenced in addition to those monitored by OBERAXE (women, LGTBIQ+ individuals, etc.). In the majority of notifications (94.5%), this has not occurred, except in 5% of cases where the discourse is also directed against women and 0.28% against LGTBIQ+ people.

Figure 7 illustrates the most prevalent types of hate speech on each respective platform. YouTube and TikTok are dominated by content that incites violence through direct or indirect threats, with 39% and 29% respectively. Meanwhile, messages that dehumanise target groups are

Graph 6. **Distribution of hate speech types (%) in each target group**

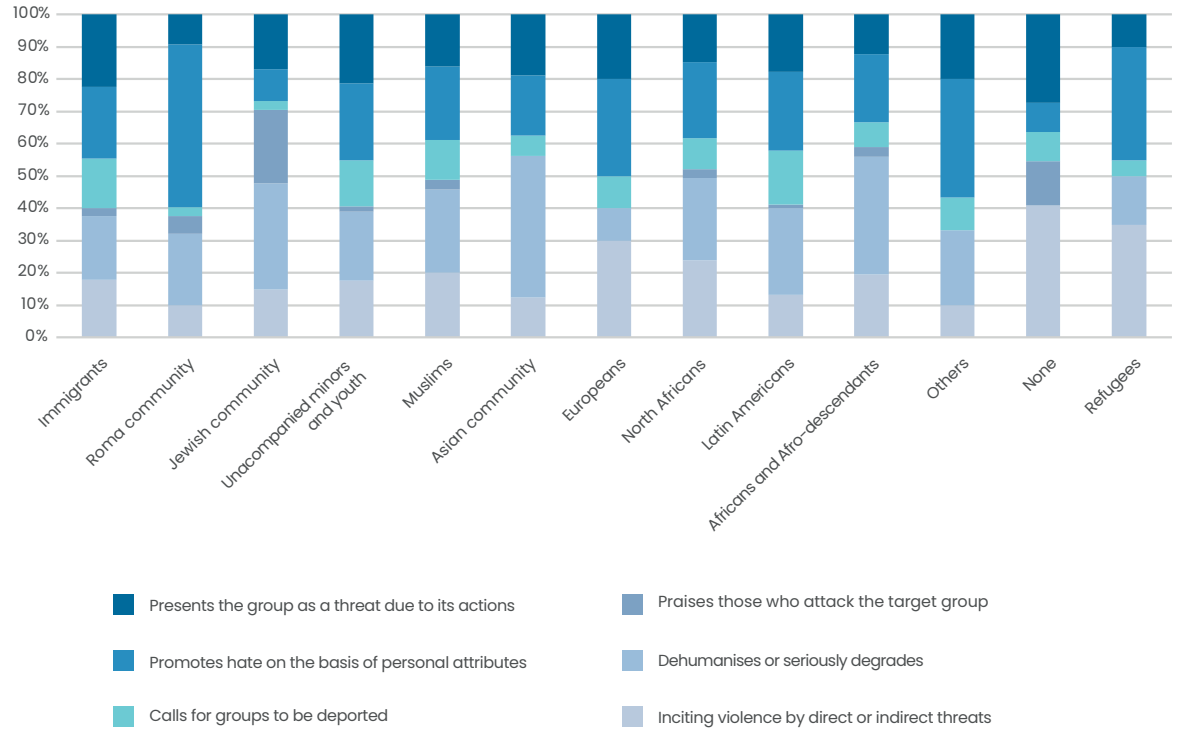
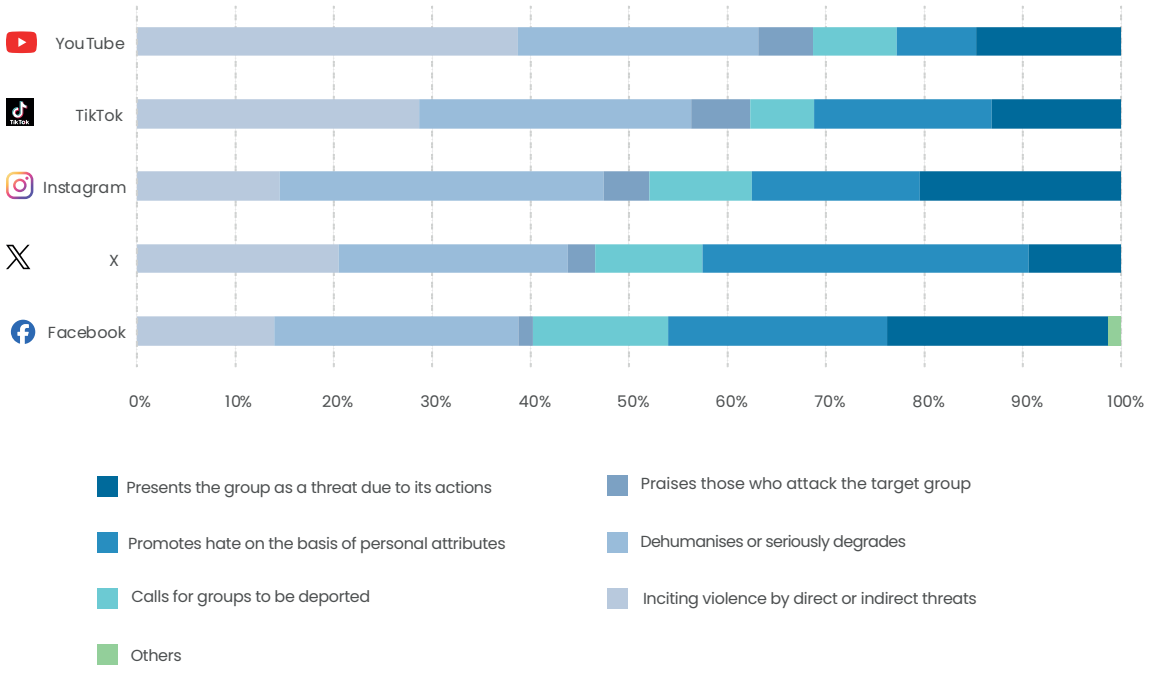


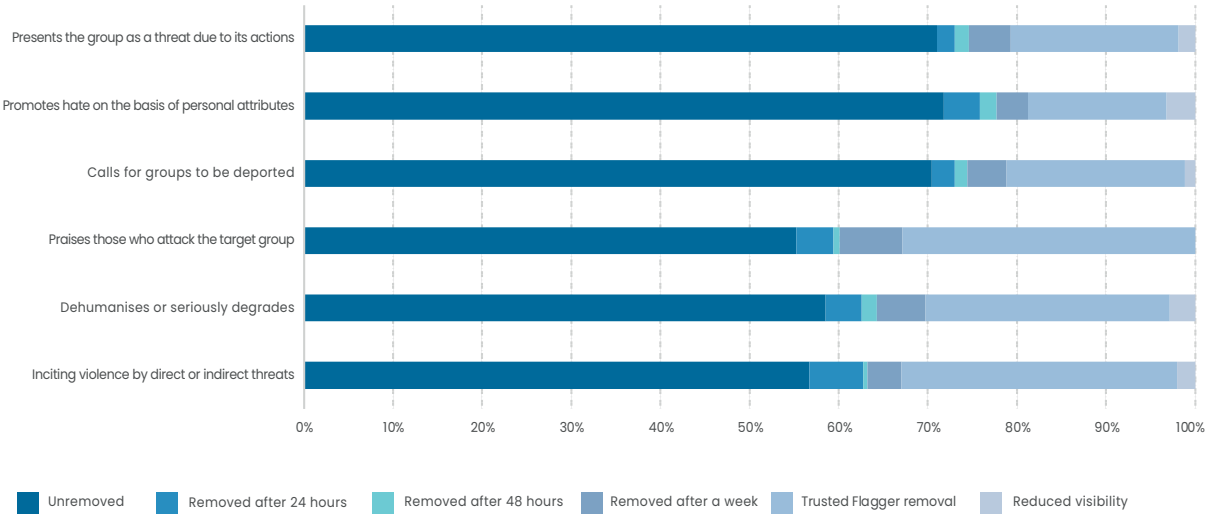
Table 3. **Distribution of hate speech types (%) in each target group. Note: the percentages of the rows may add up to more than 100 as the same content may correspond to several typologies.**

	Inciting violence with threats	Dehumanises or seriously degrades	Praises those who attack the the target group	Incites expulsion from the collective	Promotes discrediting	Presents the group as a threat
Immigrants	26%	29%	4%	22%	32%	32%
Roma community	13%	28%	7%	3%	63%	11%
Jewish community	21%	47%	32%	4%	14%	24%
Unaccompanied children and adolescents	27%	32%	3%	22%	36%	32%
Muslims	30%	38%	4%	18%	34%	24%
Asian Community	17%	58%	0%	8%	25%	25%
Europeans	33%	11%	0%	11%	33%	22%
North Africa	34%	37%	4%	14%	34%	21%
Latin Americans	19%	38%	2%	23%	34%	25%
Africans and people of African descent	26%	48%	4%	10%	28%	16%
Other groups	17%	39%	0%	17%	61%	33%
No group	53%	0%	18%	12%	12%	35%
Refugees	47%	20%	0%	7%	47%	13%

Graph 7. **Distribution of hate speech types by social network**



Graph 8. **Types of hate speech according to platforms' reaction to removal**



more frequent on Instagram (33%) and Facebook (25%), where there is also a greater presence of content that presents the target group as a threat to citizens. In the case of X, the most recurrent type of discourse is that which discredits on the basis of personal attributes, with a prevalence of 33%.

A typology of hate speech is presented in Figure 8, which illustrates the different reactions of the social media platforms to the removal of hate speech content. It is noted that the majority of posts inciting violence with direct or indirect threats (57%) were not removed, while only 6% were removed within 24 hours. Similarly, 59% of content that dehumanises or seriously degrades groups has not been removed, and only 4% was removed on the first day. Notably, posts praising those attacking the target group had a 55% non-removal rate, and 33% of these were via the trusted flagger route. In contrast, content inciting expulsion of the collective and content promoting hate based on personal attributes showed a higher proportion of non-removal, with 70% and 72% respectively, reflecting less immediate action against these forms of hate speech.

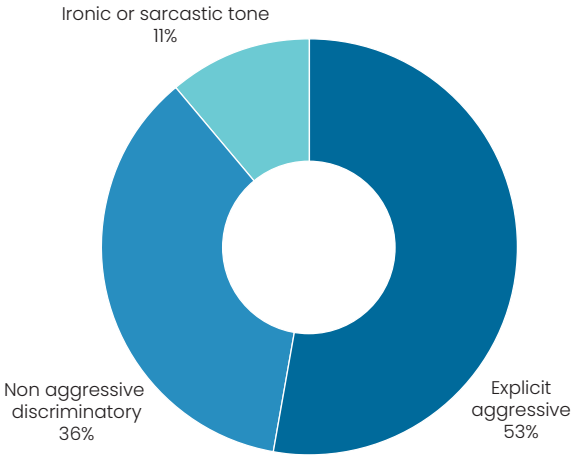
4.2.3. Expression of Hate Speech

Of the three categories considered for the expression of hate speech (Graph 9), the results show that explicit aggressive speech is the most frequent, appearing in 53% of the reported content. Examples of monitored content*: "The only way is to come together and start a hunt against these rats..."; "These scum have to be eliminated".* Non-aggressive discriminatory discourse is observed in 36% of the monitored content with examples such as the following: "*All foreigners must be thrown out before this gets out of hand"; or "Immigration... What could go wrong?"* And the ironic or sarcastic tone is present in 11% of the content. Some of the contents monitored are: "*another monkey that escaped from the zoo"; "if that Austrian painter were still around, you'd be coming out of a chimney"*. Irony and sarcasm have increased by four percentage points in the last year. This increase is evidence of a growing complexity in the communicative strategies used to disseminate discriminatory messages. These expressions, disguised as ambiguous and culturally coded as humour or criticism, make it difficult for them to be socially recognised as hate speech, which favours their normalisation and reproduction in digital spaces.

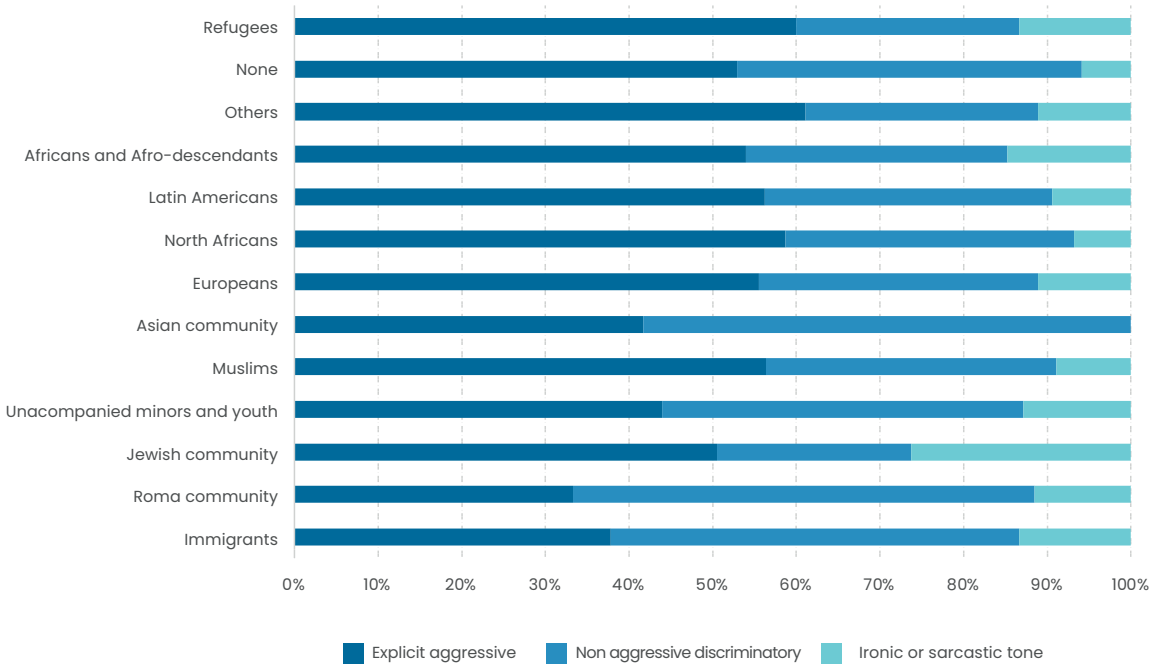
With regard to the expression of hate speech as observed among the target group, some differences can be discerned. However, in the majority of cases, the use of explicit and aggressive speech is predominant. This type of discourse occurs in 61% of cases directed at other groups, 60% towards refugees, 59% towards North Africans and 56% in the context of Islamophobia. However, Figure 10 reveals that in the case of hate speech directed at the Asian community (58%), the Roma community (55%), and immigrants (49%), there is a higher prevalence of non-aggressive discriminatory speech.

Graph 11 illustrates the types of hate speech employed on the various platforms, indicating their prevalence. A content analysis of X and Instagram reveals that 72% and 62% of the reported content, respectively, is of an explicit and aggressive nature. In contrast, non-aggressive discriminatory speech is more frequent on YouTube

Graph 9. **Frequency of the expression of hate speech**



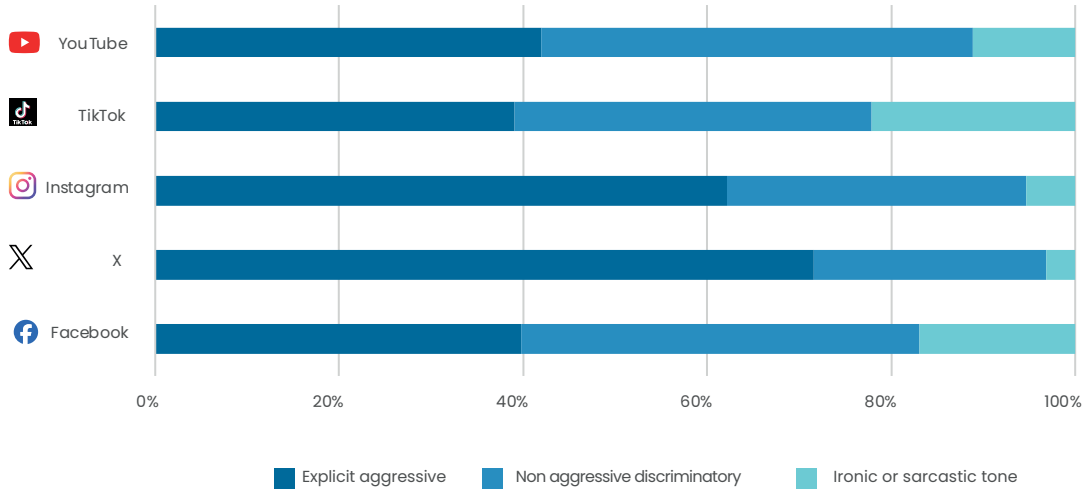
Graph 10. **Distribution of hate speech expression according to target group**



(47%) and Facebook (43%). In contrast, the ironic tone is most prevalent on TikTok, where it accounts for 22% of reported cases.

The complexity of identifying hate speech when an ironic or sarcastic tone is used is evident in the data on content removal by platform. Of the 2,870 notifications made, 323 related to ironic content, 44% of which was removed. In comparison, content with aggressive speech was removed 38% of the time, while non-aggressive content had a removal rate of 29%.

Graph 11. **Distribution of type of hate speech expression by internet platform**



4.2.4. Prototypical Episode Linked to Hate Speech

Hostility on social media towards people of foreign origin is no longer a one-off phenomenon, linked exclusively to specific events. In 2024, this trend has established itself as a persistent manifestation, articulated around stereotypes, hoaxes and polarisation that find in social media a fast way to spread. While some peaks of activity respond to specific events, a significant proportion of hate speech remains active even in the absence of events, which is evidence of a structural basis of racism and xenophobia. In this context, a series of prototypical episodes can be observed that act as recurrent triggers of hate speech, the most predominant being the link between immigration and public safety.

34% of racist and xenophobic comments and/or images are related to describing people of foreign origin as prone to engaging in violence and/or theft. The link to public safety remains latent in 2024, being the most predominant incident in the content communicated to the platforms. A considerable part of this content promotes the idea that immigration is synonymous with threat, appealing to the need for "secure borders" in the face of a supposed "migrant invasion", and even advocating mass deportation of certain target groups, especially people from North Africa.

It should be pointed out that around 40% of the content referring to public safety is only a perception and is not based on true, current events that have occurred in Spain, and therefore may refers to false information, hoaxes or decontextualised incidents.

An example of this are the hoaxes and/or fake news that spread rapidly on social media, with the aim of characterising people of foreign origin as a threat to public safety. A prominent case was the August murder of a child in Mocejón (Toledo), which generated social alarm and falsely promoted the perception of migrants, especially those from North Africa, as a threat. This episode also contributed to social fragmentation, in a context marked by the debate on migration policies and the management of reception centres for foreign minors.

Another important example was that of the DANA catastrophe of 29 October, which particularly affected the province of Valencia, but also Castilla-La Mancha and Andalusia. In this case, different target groups, such as North Africans, Muslims and the Roma community, were criminalised and linked to looting and theft. This approach stigmatised these groups as threats to citizens at a time of grief, vulnerability, uncertainty and emotional toll for the population in the affected areas. Subsequently, in the weeks marked by citizen solidarity, a hate speech narrative, based on hoaxes, spread that people of immigrant origin, and particularly Muslim women, were not providing any help in the affected towns, while allegedly taking advantage of state subsidies and Spanish citizens' taxes, thereby fuelling hostility towards immigrants.

But it is not only events in Spain that trigger these discourses. Events outside Spain, such as the "terrorist attack" in Magdeburg (Germany), also trigger hate speech comments, which are linked to the perception of public danger. These messages seek to generalise fear and mistrust towards certain groups, promoting fear in the population.

However, a significant proportion of the monitored hate speech, 21%, does not respond to any particular prototypical episode. This indicates that hostility towards migrants occurs autonomously, independently of specific social events. This suggests a structural basis of prejudice and discriminatory attitudes that are constantly active, fuelled by

misinformation, rumours and conspiratorial content. This suggests that anti-immigrant discourse spread through social media —mainly via fake news— is also driven by stereotypes about people of foreign origin.

On the other hand, public policies continue to be a frequent target of hate speech, which exploits administrative decisions to fuel theories of institutional privilege towards migrants. 9% of all monitored content in 2024 is directly linked to this prototypical episode. In February, coinciding with the news of the transfer of more than a thousand immigrants from the Canary Islands to the mainland, there was a 19% increase in the number of cases, with posts such as "*Terrible news. They bring nothing but disease and misfortune*". Another peak was observed in April, when new arrivals generated 16% of hate speech related to migration policies. The trend shows a correlation between inclusion-oriented decisions by public administrations and social responses marked by hostility and misinformation.

The discourse linked to "public policy" shows a variable trend. There is a fluctuation over the months, with periods of lows and sporadic peaks that possibly reflect the racist discourse associated with the activity in reception centres for immigrants, as well as the management of migration policy and the situation in the Canary Islands with regard to the reception of unaccompanied children and young people, together with the transfer to centres in other autonomous communities.

Added to this narrative is the content linked to the arrival of boats on the Canary Islands' coasts, with particular intensity from June onwards, which generated a new wave of hate speech. This episode accounted for 7% of the content recorded throughout the year, with 14% of the total in June following the news of the rescue of 10 boats with more than 500 people on board. The most repeated messages were of extreme symbolic violence, with expressions such as "*What does the navy have missiles for?*" or *"Cement and to the bottom of the sea"*. Such comments, in addition to trivialising death and human suffering, normalise discourses of extermination which, although illegal, some circulate with impunity on digital platforms.

Likewise, sport, which accounted for 4% of the content monitored in 2024, was also an arena where racism and xenophobia were intensely expressed. Throughout the year, significant peaks of hate speech linked to sporting events have been identified, with a particularly high incidence in football. In this context, racism has mainly been directed at players, as in the case of Vinícius Júnior, whose episode of discrimination brought racism back to the centre of the debate both in stadiums and on digital platforms. This phenomenon reached its peak in March, with 18% of the content related to this specific case. During this period, insults, mockery and questioning of identity were recurrent, as exemplified by expressions such as *"Monicius in pure form "*.

The impact of hate speech in sport is not only limited to comments on social media. There have also been violent incidents and intolerant demonstrations in stadiums. A clear example occurred on 7 November during a Europa League match between Ajax Amsterdam and Maccabi Haifa of Israel, where xenophobic chanting and physical aggression were reported among fans. The hate content identified in this case targeted both Jewish and Muslim people.

Moreover, the phenomenon of hate speech is not only confined to football. At the Paris Olympics in July and August, which coincided with the participation of teams such as Morocco, the narrative of hate became more visible. There was also a significant spike in the UEFA Super Cup, where there was a 29% increase in sports-related hate messages. These messages included discriminatory comments towards athletes of foreign origin representing Spain, under questions of identity and also towards a player of the Moroccan national team who received a wave of racist messages such as "*Even if he was born in he is still a 💩 Arab*". These comments reflect hostility towards the identity of athletes who, despite representing their country, are perceived as "foreign" by some sectors of society, highlighting the racism and xenophobia underlying much of this discourse.

Hate speech in sport is not only limited to the verbalisation of racist or xenophobic insults, but also creates barriers that hinder the integration of people of foreign origin.

There were also significant peaks of Islamophobic content, coinciding with symbolic dates in the Muslim calendar or following attacks in other European countries. Much of this discourse consisted of widespread accusations that Islam is incompatible with Western democratic values. Muslim women, especially those wearing hijabs, were the main victims of these attacks, being presented as symbols of an alleged "cultural oppression" or even as threats to the "freedom of Spanish women". Islamophobic discourse stigmatises, generates fear and conditions the daily lives of thousands of people who profess Islam in Spain.

Likewise, anti-Roma sentiment is manifested through videos of neighbourhood conflicts, which are used to collectively criminalise the Roma community, reinforcing prejudices about their alleged links to crime, economic irregularity or violence. This type of content, loaded with derogatory language and discriminatory humour, continues to fuel a negative

perception that has a direct impact on the social integration of the most discriminated ethnic minority in Spain.

Anti-Semitism also resurfaced, particularly in the wake of the Gaza conflict. Numerous publications with Holocaust denial messages, caricatures and Nazi symbolism were detected. Also, some publications contained a narrative that held the Jewish community collectively responsible for all violent events in the Middle East. This type of discourse made no distinction between the Jewish community as a whole and the actions of the Israeli government, perpetuating the idea that Jewish people, as a whole, are responsible for violent conflict.

Overall, the data show that hate speech is not limited to moments of crisis or conflict, but finds in various episodes, whether real, fictional or symbolic, a catalyst for its propagation. Social media are consolidating as a space where hostility towards people of foreign origin is naturalised.

Graph 12 illustrates the proportion of content reported to the relevant platforms in relation to the prototypical episode that prompted its submission or to which it is linked.

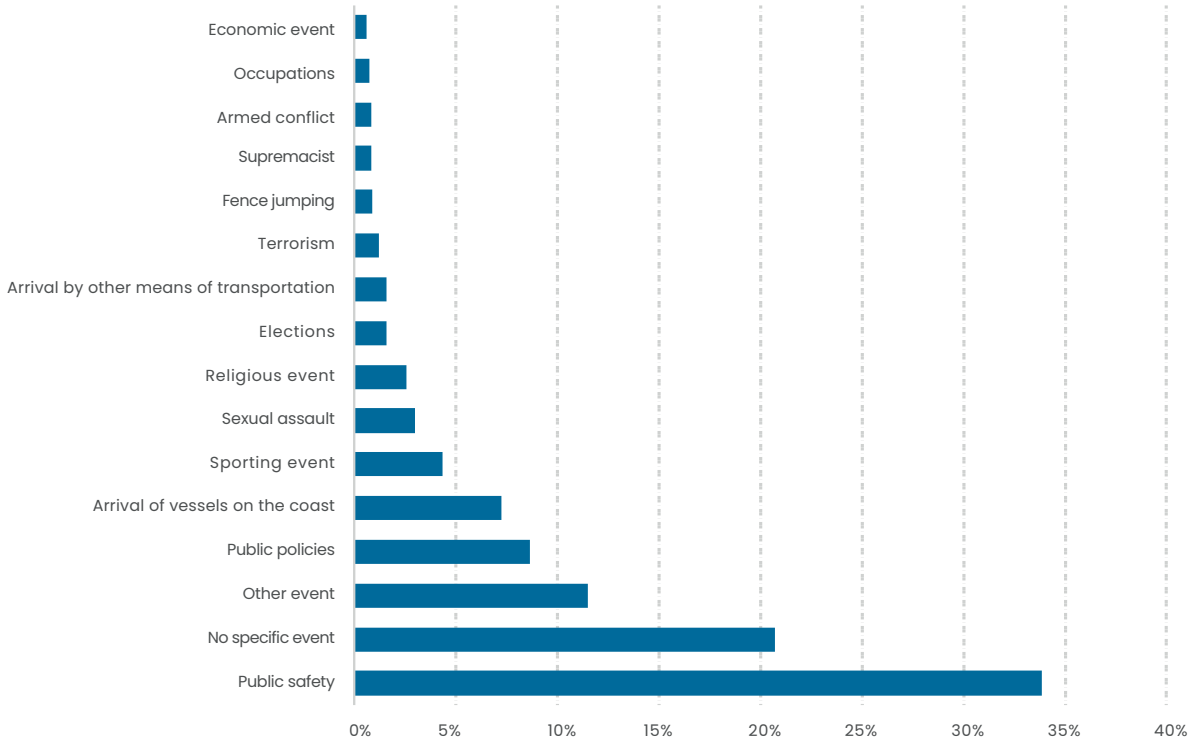
2024's analysis confirms that hate speech on social media is not only the result of moments of crisis, but a structural phenomenon. The constant dissemination of hatred, in the form of comments, memes or fake news, not only affects migrants, but also undermines the fundamental principles of a democratic society.

The evolution of hate speech over the year is presented below, based on the main prototypical episodes identified during the year (Graph 13). The data show that spikes in hostile discourse are closely linked to easily recognisable conjunctural events.

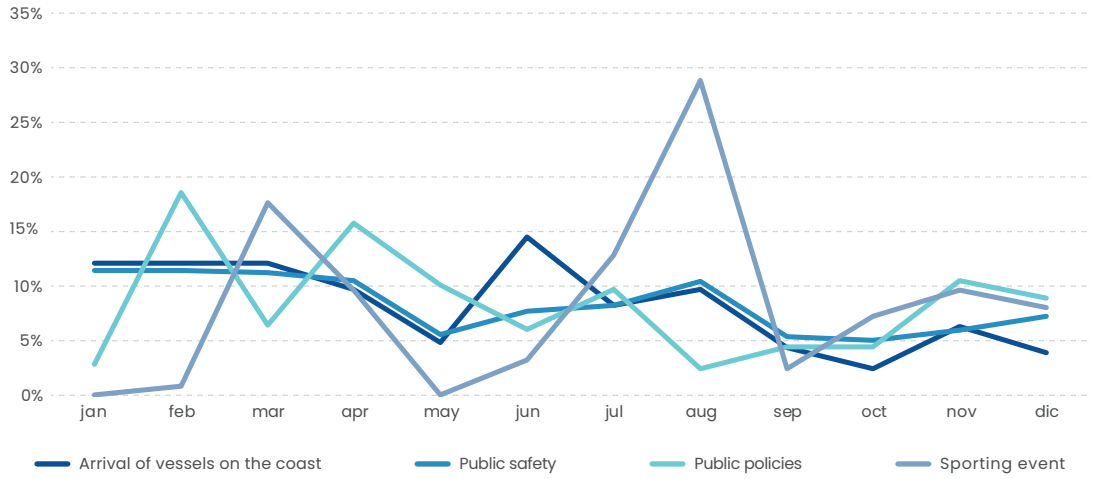
The arrival of vessels on the coasts acted as a constant trigger, with rates above 12% in the first three months of the year and a spike in June (14%). This narrative reinforces the idea of an external threat and is recurrently used to legitimise discourses of rejection towards migrants and refugees.

Likewise, public safety was a structural narrative in hate speech, with a sustained and notorious presence in the first months of the year, reaching its peak in January

Graph 12. **Distribution of hate speech reported to platforms by link to a prototypical episode**



Graph 13. **Evolution of the main prototypical episodes in 2024**



and February (both with 11%), associated with the criminalisation of certain groups.

Meanwhile, public policy took centre stage in February (19%) and April (16%), coinciding with institutional debates that provoked strong reactions on social media. These debates generated a marked rejection of populations perceived as beneficiaries of aid or inclusion measures, especially migrants and minorities, becoming a recurrent focus of hostile discourse.

Finally, sporting events emerged as an uncommon but highly impactful trigger, reaching a notable peak in August (29%), coinciding with international competitions (e.g., the Paris Olympic Games). In these contexts, there was evidence of an increase in xenophobic expressions linked to the performance of athletes of foreign origin or the representation of specific countries, reflecting how sport can function as a factor in excluding national identities.

Figure 14 shows the prototypical episodes according to the target population group most affected. A lack of public safety is the prototypical narrative that predominates across all target groups, although it is most prominent in discourse directed at people from North Africa, with 51% of communications related to this group linked to this particular narrative. The same is true for Latin Americans (48%), and unaccompanied children and adolescents (47%).

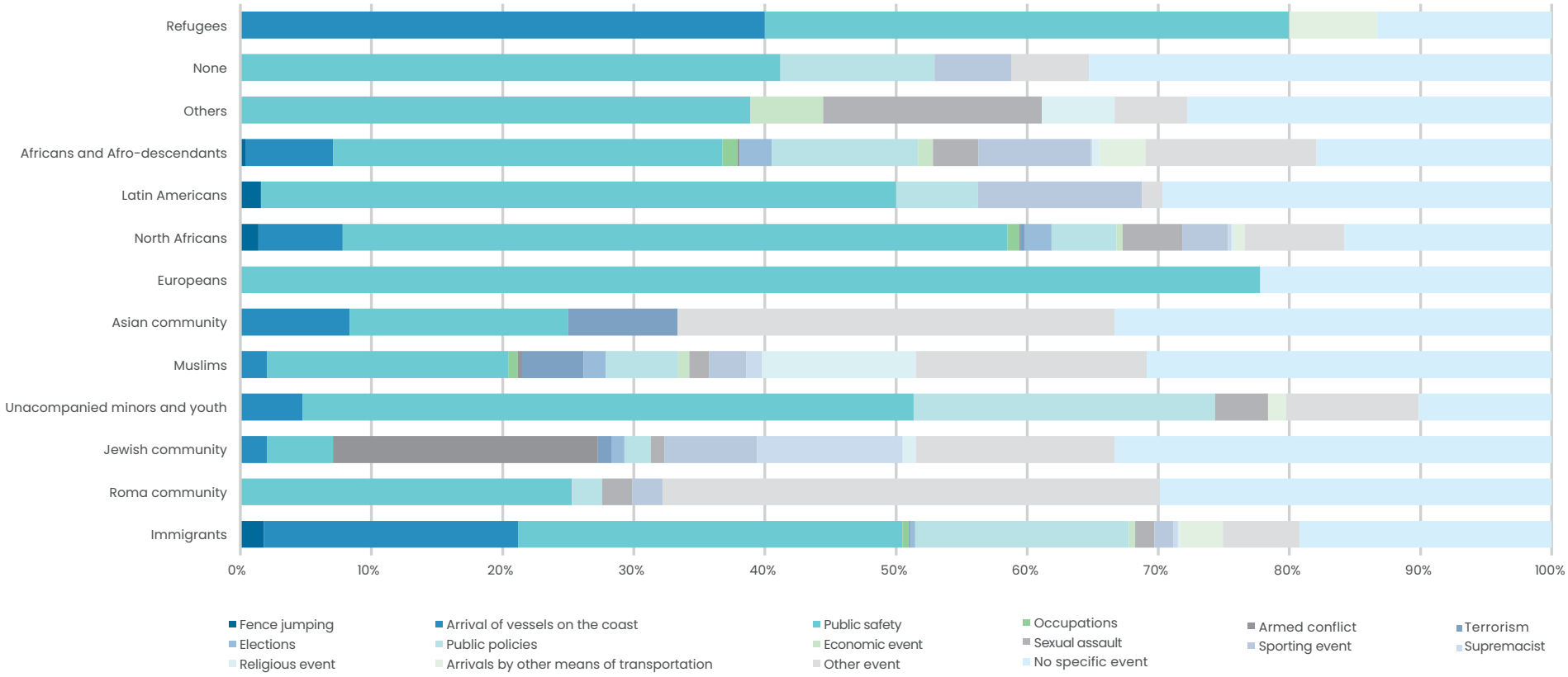
In terms of the prototypical public policy episode, 23% of the cases are related to unaccompanied minor children and 16% to immigration.

On the other hand, the prototypical episode of boat arrivals on the coasts is linked in 19% of cases to migrants.

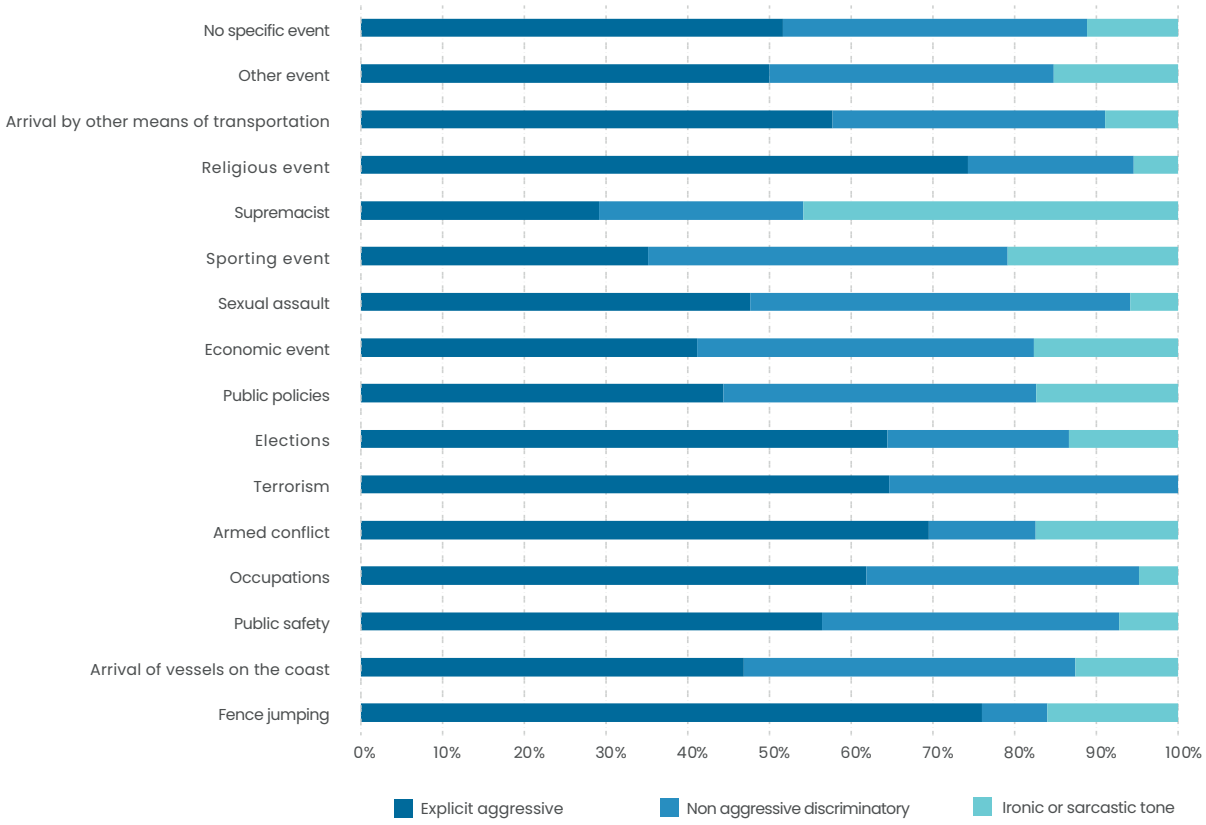
Graph 15 shows the relationship between the episodes that give rise to hate speech and the nature of the language used. The most frequent episode identified is that of "public safety", in which the use of explicit aggressive discourse predominates, present in 56% of the communications linked to this episode. Also, this type of language predominates in situations such as fence-jumping in Ceuta and Melilla (76%), religious events (74%) and armed conflicts (70%).

In terms of less explicit hate speech, non-aggressive discriminatory language is most frequent in contexts such as sexual assault (46%) and sporting events (44%). Furthermore, speeches with a tone of irony or sarcasm are linked to episodes of a "supremacist" nature, reaching 46% of the analysed content related to this episode.

Graph 14. **Distribution of prototypical hate speech episodes by target group**



Graph 15. **Distribution of hate speech according to prototypical episode**



4.2.5. Hate speech towards women

Hate speech on social media, as in other areas of society, reflects and reinforces the dynamics of discrimination, hatred and intolerance that are present in our society.

Individuals and groups can be targets of hate speech on different grounds or personal characteristics ("race", colour, national or ethnic origin, age, disability, religion, sex, gender identity and sexual orientation) or combinations of these. In this respect, when monitoring hate speech and analysing its trends, gender is one of the most important dimensions to take into consideration.

Analysis of the data monitored in relation to hate speech in 2024 shows how sexist, racist, xenophobic and Islamophobic comments often go hand in hand, particularly affecting certain groups of women. According to the data, although the majority of the contents analysed are related to the generic masculine (90%), 6% of the contents present a linguistic mark in feminine. This figure, although apparently low, reveals a presence of hate speech or manifestations directed specifically at women, and particularly at certain target groups.

Thus, a high percentage of female-branded content is targeted at Muslim women (60%) and women of African descent (21%).

In the case of Muslim women, in particular, double or triple discrimination is observed: because of their gender, their origin and/or religious affiliation. The comments towards these women, the derogatory terms with which they are described, are often associated with negative stereotypes about Islam, and based on entrenched gender stereotypes, perceiving them as submissive or subordinate. Examples such as "*She is mentally ill. Let's see how long she keeps smiling when she marries a Muslim and his customs"* or *"Parasols are for the summer... what a filthy disgrace"* show double or triple discrimination against these women.

A recurrent element in this type of discourse is the reference to the use of the hijab or Islamic headscarf, a symbol that sometimes becomes an object of criticism and mockery. The rejection of the hijab is not only based on prejudice towards the Muslim religion, but, far from being understood as a personal choice or a religious symbol, its use is presented as a form of oppression, thus encouraging discourse that denies the autonomy of Muslim women. This rejection is not limited

to religious criticism, but is part of a broader discourse that questions their ability to decide for themselves and reinforces sexist, racist and Islamophobic stereotypes.

Similarly, comments towards women of African descent demonstrate once again the intersection of misogyny and racism. Expressions such as *"I don't need that fucking black woman to defend me"* or *"In Asturias, they grow cotton and someone has to pick it"* exemplify how hatred towards women of African descent is built on a racist and sexist narrative, reducing them to traditionally stereotyped roles. Such comments are not only evidence of racism, but also of persistent sexism that affects women.

When analysing the relationship of the content with the prototypical episodes that give rise to this type of discourse, we found that 32% of the comments with a feminine linguistic mark are not linked to any specific event. This is particularly relevant, as it reflects the fact that hatred towards women is constantly reproduced, feeding on entrenched social attitudes, stereotypes and prejudices. However, it is also observed that 13% of the contents are related to public safety, which shows how collective fear of people or groups that are perceived as different or social violence is projected onto women, perceiving them as a threat to social coexistence or the cultural identity of the majority group.

Comments such as *"Women are threatened by the riff-raff, enemies of ham"* or *"Españistan (Spainistan). The Muslim feminist terrorist dictatorship of violence. Evil and death"* are examples of how hate speech uses fear of what is different and what is "non-Western" to justify violence against women. These comments incite xenophobia and racism, and also fuel social concerns about cultural and social changes, portraying such women as a threat to citizenship.

In this context, an intersectional view is indispensable to understand the complexity of hate speech towards these groups, and how it is articulated in multiple layers that combine sexism, racism, Islamophobia and xenophobia. Women of foreign origin, women of African descent or Muslim women in Spain are not only hated because of their gender, but also because they belong to socially stigmatised groups, which places them in a particularly vulnerable situation.

5. CONCLUSIONS

The analysis of hate speech on social media throughout 2024 shows it to be a structural and persistent phenomenon, which transcends circumstantial events to become a constant element on digital platforms. Although certain peaks of activity are linked to specific events such as migrant movement, natural disasters or sporting events, a significant proportion of hostile messages persist regardless of any particular event. This continuity reveals a solid base of racism, xenophobia, Islamophobia, anti-Roma sentiment and anti-Semitism, which finds in social media a space for its reproduction and normalisation.

The data obtained show that the five monitored platforms (X, Facebook, Instagram, TikTok and YouTube) removed only 36% of the content reported to them in 2024, despite it potentially constituting a criminal offence, an administrative violation or a breach of their own community guidelines.

In terms of target groups, people of North African origin, followed by people of African descent, immigrants and Muslims are the four target categories to which most hate content is directed (35%, 24%, 21% and 20% respectively of the total monitored). It is also noteworthy that 5% of the content is directed towards unaccompanied children and young people, albeit at a lower percentage.

One of the key findings of the report is that public safety, with 34% of the reported content, continues to be the prototypical episode most linked to hate speech, especially towards North Africans and unaccompanied children and adolescents. This narrative, which associates immigration with crime, has been fuelled by disinformation, hoaxes and manipulated news, most of which do not correspond to real or recent events, and which in 40% of cases consist of false or decontextualized content. The constant appeal to collective fear, underpinned by stereotypes of dangerousness, facilitates the justification of repressive measures such as deportation or border closures, undermining democratic principles and fundamental human rights.

Moreover, the instrumentalisation of specific episodes such as the murder in Mocejón or the DANA catastrophe in Valencia demonstrates how certain crises can be exploited to fuel hate speech. In these cases, migrants are not only criminalised, but a perception of existential threat is fuelled, which fragments social cohesion and generates a climate of polarisation and intolerance. The narrative that questions the legitimacy of foreigners to receive aid or participate in solidarity processes further accentuates this symbolic exclusion, presenting them as undeserving beneficiaries rather than full citizens.

It is also noted that hate speech is not limited to events in Spain. International incidents also trigger waves of hostile comments on social media in Spain. This phenomenon demonstrates the existence of a globalised narrative of the migrant or Islamic threat, where stigmatised groups are treated as a homogenous bloc, and any incident is used as a justification to reinforce collective rejection.

Another of the prototypical episodes that generate hate speech is that linked to public policy (9%). Public policies, such as the transfer of people from the Canary Islands to the mainland, provoke negative reactions expressing strong opposition to the inclusion of migrants in the social fabric. These narratives, which denounce alleged institutional privileges, generate a perception of injustice among the indigenous population and erode trust in democratic institutions. This type of discourse has shown a fluctuating pattern, with peaks coinciding with specific administrative decisions, revealing a social sensitivity to migration issues and diversity management.

The sports sphere, with 4% of the reported content, is also a space where hate speech, particularly racist and xenophobic, is reproduced and amplified. Cases of football players highlight how racism is openly expressed both in stadiums and on social media. Hostility directed towards athletes of foreign origin, even when they represent Spain in international competitions, reveals a deep conflict over national identity and belonging. This type of discourse does not only affect individual athletes, but has a symbolic effect of exclusion towards entire communities by denying their full integration into society.

Special mention should be made of the impact of hate speech directed at women (5%), particularly those at the intersection of multiple forms of discrimination, such as Muslim women and women of African descent. The combination of sexism, racism and Islamophobia gives rise to a specific symbolic violence that reinforces colonial and patriarchal stereotypes, and is expressed through mockery, scorn or questioning of their ability to decide about their own lives. Women who wear the hijab, for example, are recurrent targets of attacks that not only deny their autonomy, but also portray them as cultural threats.

All of this reinforces the need to apply an intersectional approach to the analysis of hate speech, understanding that the different axes of discrimination (race, gender, religion, class) do not act in isolation, but intertwine and mutually reinforce each other, generating situations of extreme vulnerability.

In terms of the language used, the analysis reveals that the most explicit and aggressive discourses predominate in contexts such as public safety, fence jumping and religious or armed conflicts. By contrast, in episodes such as sporting events or sexual assaults, the language tends to be more subtle, often ironic or sarcastic. This variability in tone demonstrates that hate speech is not always presented in overt or reportable forms, making it difficult for digital platforms to detect and sanction. However, even less explicit comments have a strong symbolic impact, as they contribute to naturalising prejudices and creating a hostile social climate towards certain groups.

In terms of the type of content, there is a pattern of messages and/or images that dehumanise or seriously degrade the people they are aimed at (39%), inciting violence with direct or indirect threats (29%), and inciting the expulsion of the immigrant community in 15% of the content monitored. The process of stigmatisation ultimately results in the creation of a hostile environment. It can also fuel fear and resentment toward certain population groups, which may lead to greater fragmentation and social conflict.

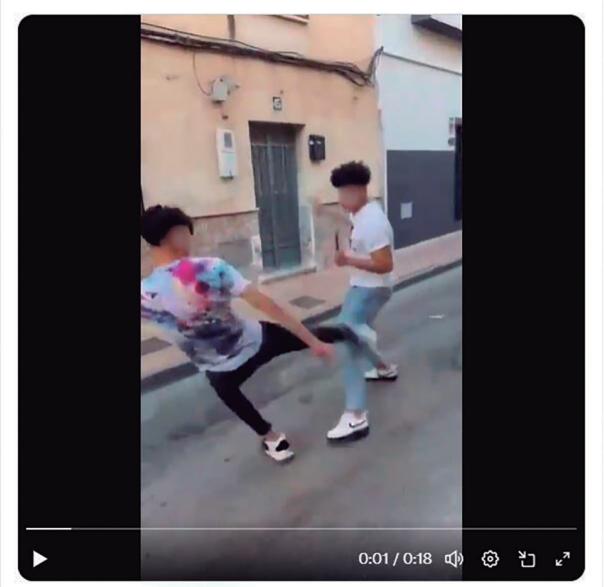
Overall, the data show that hate speech on social media in Spain is not a one-off phenomenon. It is a form of structural symbolic violence that is sustained by the reproduction of stereotypes, misinformation, and political and social polarisation. The impact of this discrimination and intolerance goes beyond the individual harm it can cause to those directly affected. It has serious consequences for social cohesion, democratic functioning and human rights, eroding mutual trust, fuelling fear of the other and undermining the values of pluralism and inclusion.

It is therefore necessary to address hate speech directed towards people of foreign origin as a multifaceted phenomenon that requires coordinated responses from public institutions, digital platforms, the media and civil society. Actions must go beyond content moderation and must include awareness-raising and the promotion of alternative narratives that counteract the dehumanisation of target groups in order to build a more inclusive and resilient society in the face of hate.

6. Annex I: Examples of Hate Speech

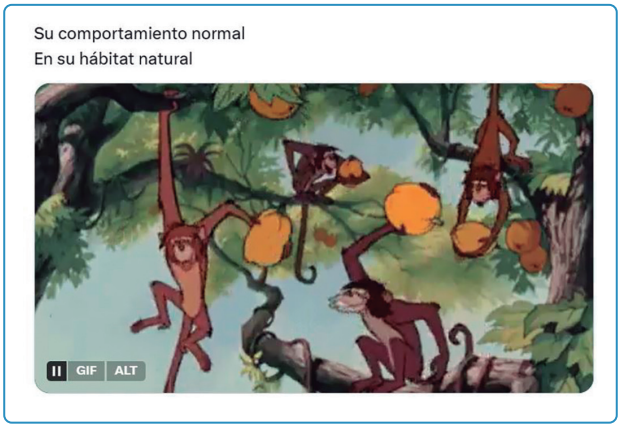
The following section presents a series of examples of illegal online hate speech content that were monitored throughout the course of the year 2024.

1. Prototypical episode: public safety and public policies



* The content uses a video of street violence to baselessly attribute criminal behaviour to target groups and presents them as a social and economic threat.

2. Hate speech against Africans and people of African descent



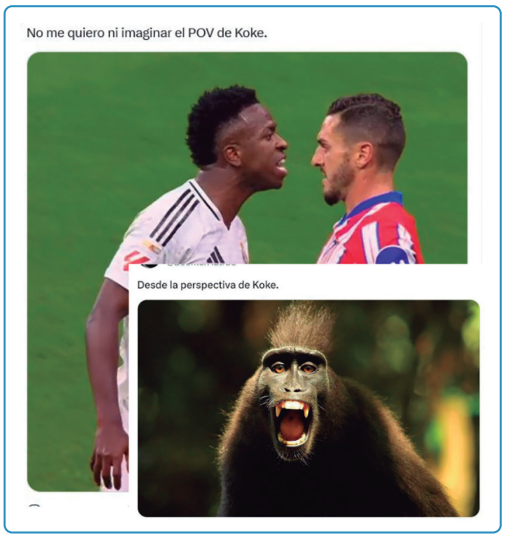
* Caption: "Their normal behaviour. In their natural habitat".

3. Hate speech against North Africans

Moro de mierda a ver si te dan una paliza y te dejan seco

* Caption: "You fucking Moor, let's see if they beat you up and leave you dry".

4. Prototypical Episode: Sporting Event

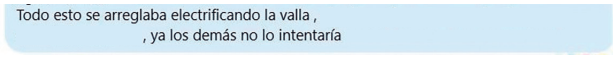


* Caption: "I don't want to imagine the Koke's POV", "From Koke's perspective". Dehumanisation of footballer Vinicius Jr at several La Liga games in 2024.

5. Prototypical Episode: Religious Event

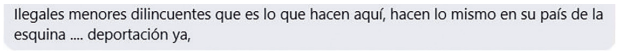
* Caption: "Moors, be careful with the ham during Ramadan. Don't be cannibals".

6. Prototypical episode: Fence jumping



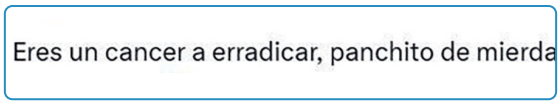
* Caption: "All this was solved by electrifying the fence. I wouldn't try anything else."

7. Hate speech against children and unaccompanied youth



* Caption: "Illegal immigrant young offenders, what are you doing here? You do the same thing in your country on the corner... Deportation now!"

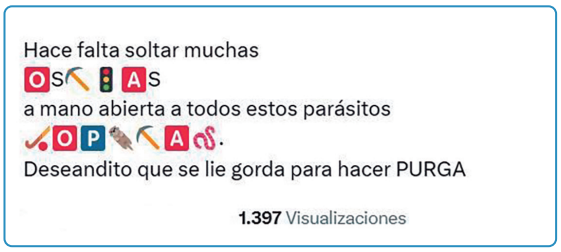
8. Hate Speech against Latin Americans



*Caption: "You are a cancer that must be eradicated, panchito shit". "Panchito" is a derogatory way of referring to a person of Latin American origin

9. Message constructed with emojis/coded language

*Mouse emojis are used to dehumanise the target group.

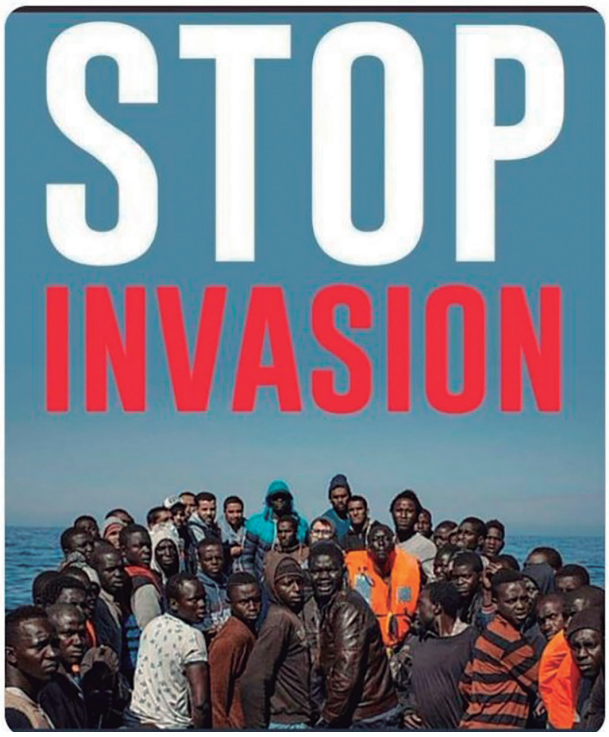


* Caption: "We need to give all these parasitic bastards a good beating with our bare hands. I hope all breaks loose so we can purge them." The message constitutes incitement to violence against this group. Emojis are combined with letters to formulate violent messages.

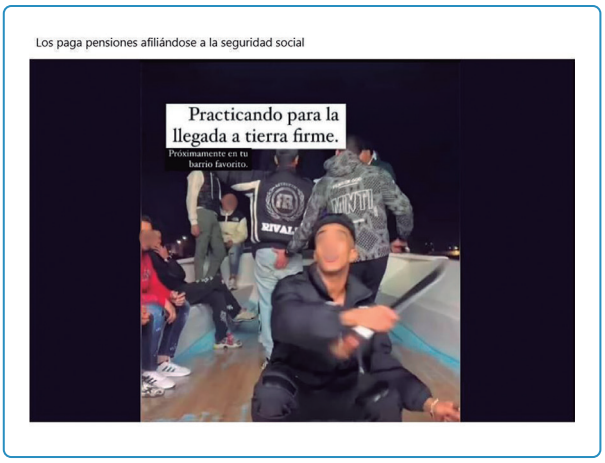
10. Anti-Semitism

* Caption: "Big-nosed rats in the run". The terms "rats" and "big-nosed" are used as a reference to the Holocaust and as a way of dehumanising and degrading Jewish people. This image is related to the New York tunnel event

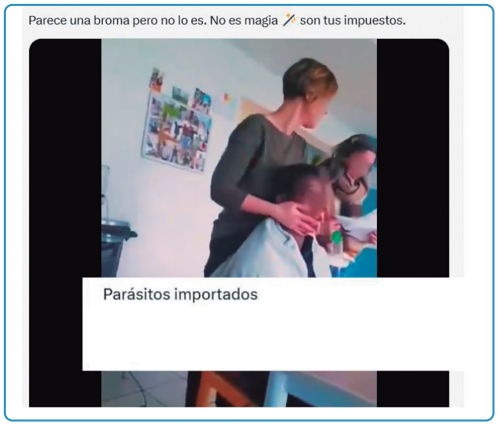
11. Prototypical Episode: Arrival of Small Boats



12. Prototypical Episode: Public Policy



* Caption: "Those who pays for the retirement pension (paga-pensiones) registering in the social security system", "Practising for the arrival on dry land".



* Caption: "Imported parasites". This publication uses misinformation to generate rejection of unaccompanied migrant minors, presenting them as unfair beneficiaries of public resources.

* "Paguitas" refers to payments from the governments, such as the Minimal Income revenue (IMV). In this case, Muslim women are pictured as beneficiaries of support from government agencies

13. Discredit is promoted on the basis of personal attributes of the group

*A narrative is used in which unaccompanied children and adolescents of Moroccan origin are presented as a threat. The "knife" is used as a means of indicating danger and the "Moroccan flag" is used to indicate origin.

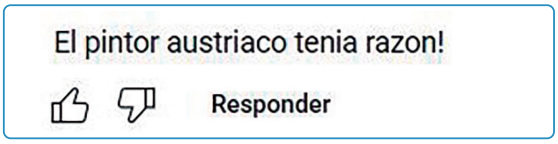
*

Caption: "Next, in the Canary Islands, the future of Spain (...) in the best cinemas" The term BroccoliHair is used for people of North Africa.

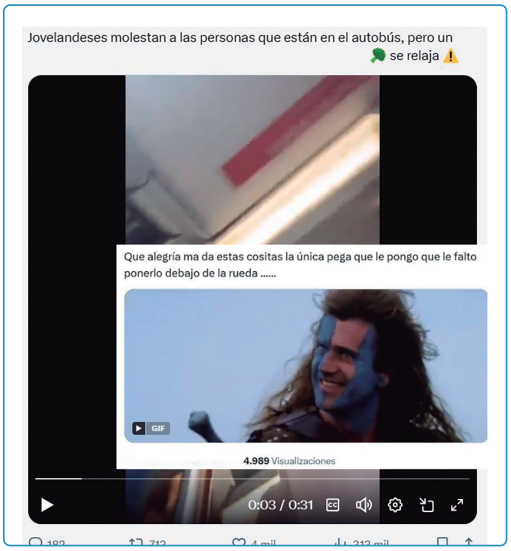
14. There are calls for this group to be deported

* Caption: "This Mohamed shit, to his country".

15. There is praise for those who attack the target group



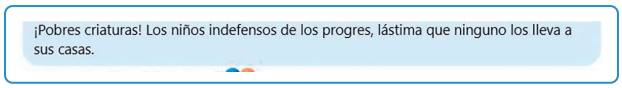
*Caption: "The Austrian painter was right". The Austrian painter is a reference to Hitler



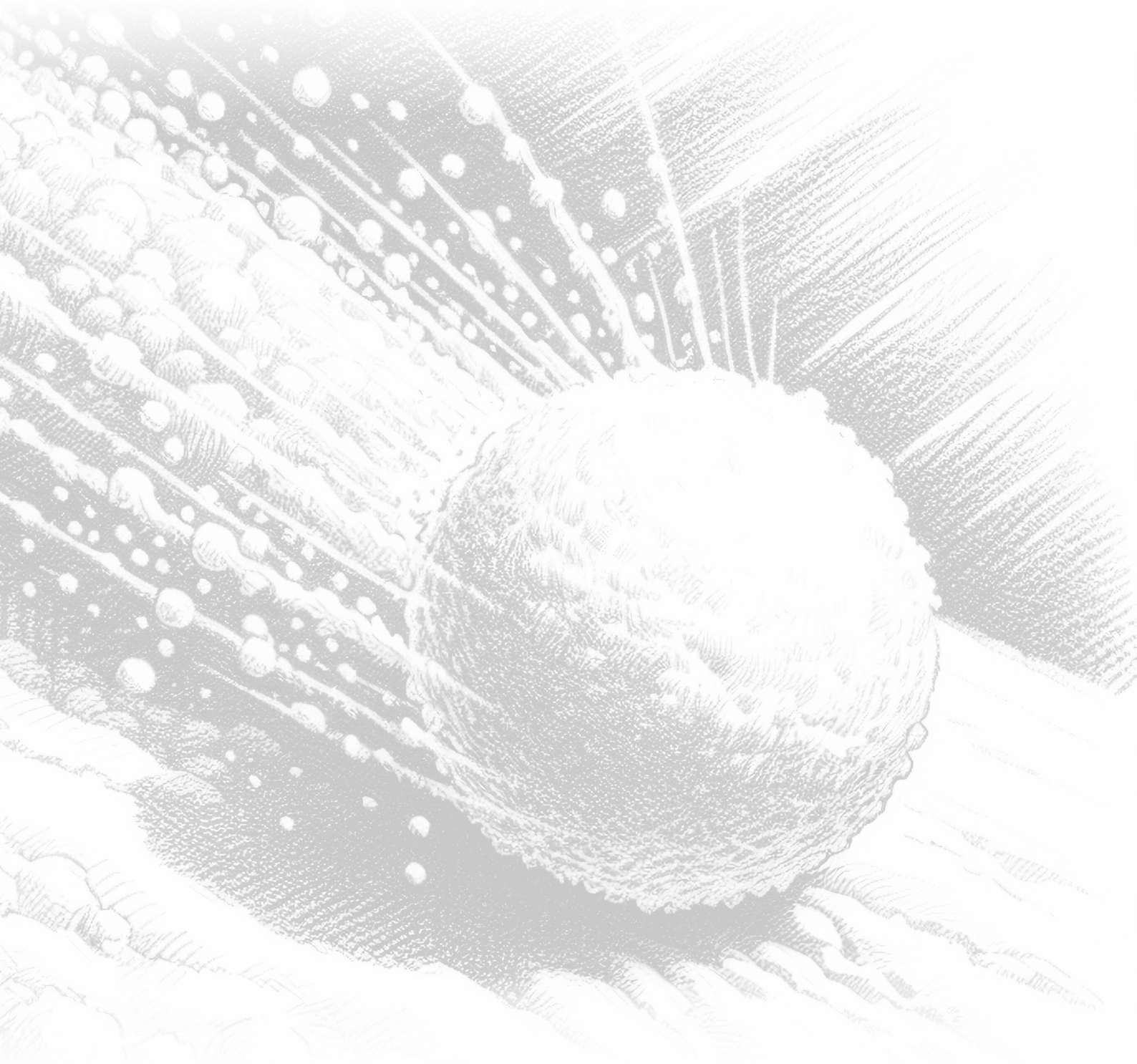
*This publication celebrates one young man berating another—identifiable as a foreigner—on a bus, insinuating that he "relaxes" after being confronted. Furthermore, the message comes with an image suggesting that the only thing he did wrong was not "running him down", thereby normalising and glorifying violence towards migrants

16. Irony or sarcasm are used

*Caption: "This happened in Ferrol, A Coruña". "Everyone is black, what a coincidence!"



* Caption: "Poor creatures! Defenceless children says the progressives minded. It is a pity they don't take home with them". Defenceless children = foreign minors



GOBIERNO DE ESPAÑA
MINISTERIO DE INCLUSIÓN, SEGURIDAD SOCIAL Y MIGRACIONES
SECRETARÍA DE ESTADO DE MIGRACIONES

oberaxe
Observatorio Español del Racismo y la Xenofobia

Co-funded by the European Union