VISUAL REVIEW | Vol. 17, No. 2, 2025 | ISSN 2695-9631
International Visual Culture Review / Revista Internacional de Cultura Visual
https://doi.org/10.62161/revvisual.v17.5764

XENOPHOBIA ON SOCIAL MEDIA: ANALYSING HATE SPEECH MESSAGING ON THE X PLATFORM

SANTA PALELLA STRACUZZI[1], MARTA SÁNCHEZ ESPARZA[2]
[1]EAE Business School, Spain
[2]Universidad Rey Juan Carlos (URCJ), Spain

KEYWORDS

X
Hate Speech
Xenophobia
Digital Media
Social Networks
Social Representations
Immigrants

ABSTRACT

The social network X enables the propagation of hate speech. The Hatemedia Project (2021-2024) investigates hate speech within digital environments associated with Spanish news media, employing a hate monitor developed via data mining techniques. A principal outcome of this project is a library that categorises five distinct types of hate: general, political, sexual, misogynistic, and xenophobic. The present study concentrates on xenophobic messages disseminated on X between June and November 2024, underscoring the adverse depiction of immigrants. It concludes that these messages engender a stigmatising narrative, with news media making a substantial contribution to their dissemination. This raises pertinent questions regarding editorial responsibility and the efficacy of current content moderation policies.

Received: 15/ 11 / 2024
Accepted: 20/ 02 / 2025

1. Introduction

The rise of social media has significantly altered global communication, enabling the rapid exchange of information while at the same time facilitating the spread of hate speech, including xenophobic rhetoric. This form of discourse exerts a considerable influence on the populace, exacerbating prejudice and contributing to a rise in hate crimes. Furthermore, it leads to the victimisation of individuals belonging to groups targeted by such discourse, resulting in psychological harm (Aldamen, 2023).

Exposure to online hate speech can reinforce negative stereotypes and discriminatory attitudes towards certain social groups (Schäfer et al., 2024). This is exacerbated in environments with pre-existing nationalist sentiments, where social networks also act as catalysts for the coordination and amplification of xenophobic actions. In Russia, for example, increased social network penetration was correlated with increases in ethnic hate crime in cities with strong nationalist sentiments (Bursztyn et al., 2019; Bursztyn et al., 2020). On the other side of the globe, in South Africa, research has verified how social networks reflect and can predict xenophobic incidents, indicating their role in shaping public sentiment and possibly inciting violence (Raborife et al., 2024).

The legislative frameworks of various nations do not offer a consistent response to the incitement of xenophobic hate speech on social media. Furthermore, this issue is not unequivocally addressed by the platforms' own self-regulatory mechanisms, leaving numerous individuals in a precarious position (Gelashvili, 2018). The absence of robust international cooperation, coupled with the technical complexities inherent in social media platforms, impedes the development of comprehensive strategies to counter xenophobic hate speech.

Numerous initiatives have been undertaken to address this escalating problem. In 2019, the United Nations published a framework strategy and action plan concerning hate speech. In Spain, the Ministry of Inclusion, Social Security, and Immigration issued a protocol and system of indicators for the detection of hate speech on social networks in 2017. Governmental observatories, such as the Spanish Observatory on Racism and Xenophobia (OBERAXE, 2017a, 2017b), have produced dossiers documenting the evolution of racism and related forms of intolerance in Spain, alongside analyses of Spanish public attitudes towards migrant populations on social networks. This observatory conducts an annual monitoring report on hate speech in social networks, with the most recent edition highlighting a concerning rise in hate messages targeting vulnerable groups, predominantly immigrants, refugees, and specific ethnic communities (OBERAXE, 2023). Racism and xenophobia constituted a substantial proportion of all monitored messages, with social media serving as a key platform for their dissemination.

Among the findings of the report, 43.5% of hate crimes in Spain in 2022 were racist or xenophobic, with a particularly high impact on immigrant communities. The platforms most involved in the dissemination of messages linked to hate speech include Facebook, Instagram, TikTok, YouTube and X (formerly Twitter), with variations in prevalence depending on the social network (OBERAXE, 2024).

The report underlines the importance of legislative tools, such as the EU's Digital Services Act, to ensure effective moderation of hate content. In addition, the dissemination of fake news and disinformation was identified as a common mechanism for fostering hatred, which calls for technological and educational countermeasures.

This study seeks to enhance the visibility of this issue by analysing xenophobic messages published on the social network X between June and November 2024. These messages were extracted as part of the data mining process of the Hatemedia Project, with the aim of identifying the characteristics of negative representations of immigrant and foreign groups within the hate speech posted on the aforementioned social network.

2. Literature Review

2.1. Hate Speech

Hate speech encompasses statements or communications that convey a denigrating, hostile, or defamatory message targeting specific groups (Fuentes Osorio, 2024). The United Nations defines it as:

> any form of communication, whether expressed through words, writing, or behaviour, that constitutes an attack or employs derogatory or discriminatory language against a person or group based on their identity. This identity may be defined by religion, ethnicity, nationality, race, colour, descent, gender, or other characteristics. (UN, 2019)

According to the Council of Europe (1997), for speech to be considered hate speech, it must disseminate, incite, promote, or justify attitudes related to racism, xenophobia, anti-Semitism, and other forms of intolerance. Furthermore, the European Commission against Racism and Intolerance, in its General Recommendation No. 15, notes that hate can be motivated by factors such as race, colour, descent, national or ethnic origin, among other personal characteristics (2016). Official definitions of hate speech particularly highlight racist and xenophobic discrimination as the main causes of rejection and hostility. On the other hand, the Spanish Ministry of Interior went so far as to differentiate up to 11 categories of discrimination affecting vulnerable groups, with racism and xenophobia being the main ones.

Some scholars, such as Valle de Frutos (2024), differentiate between hate speech and offensive speech, the latter being considered a milder subcategory. Offensive speech is defined as any communication that conveys a derogatory message towards a person or group based on cultural factors, such as religion, ethnicity, nationality, colour, gender, or other identity factors. According to this author, the primary distinction between hate speech and offensive speech lies in their potential consequences. While hate speech may incite physical violence, leading to direct discrimination, offensive speech involves verbalised and subtle rejection, explicitly referencing elements of the culture being rejected. This constitutes a means of discrediting the group by invoking parameters of morality, without resorting to direct insults, incitement to violence, or other discriminatory acts.

In any case, this discourse entails the definition and categorisation of these groups based on characteristics such as race, religion, gender, or sexual orientation, often through the dissemination of false or misleading information (Hyewook, 2023). This discourse can manifest in oral, written, or visual forms and occurs across various contexts, including the Internet, the media, and, notably, social media, where contemporary social conversation is largely conducted. Recent research, such as Arcila-Calderón et al. (2022b) and Gautam et al. (2024), demonstrates the increasing application of artificial intelligence technologies to detect and address this discriminatory discourse online, through the training of machine learning models.

2.2. Social Networks and Their Role in the Proliferation of Xenophobic Discourse

Social networks are digital platforms and environments that facilitate social interaction, communication, and information sharing among users. Although some of these platforms, such as Facebook, were already shaping public discourse before 2010, reaching 500 million users in that year (Bruneel et al., 2013), these networks gained significant prominence following the COVID-19 pandemic. This surge in popularity was driven by the increased demand for personal contact, remote work, and the digitisation of education (Kaur, 2023). Their relevance in social interaction has led companies to adopt them as a fundamental tool for marketing, customer segmentation, and product commercialisation.

The number of users on social media platforms has multiplied in recent decades, and with it, their power and influence. In 2024, the number of social media users worldwide exceeded five billion for the first time (We are Social, 2024). This substantial user base has transformed these platforms into a vehicle for articulating social movements and shaping public discourse, as evidenced by events such as the Arab Spring and Black Lives Matter (Forrest & Wexler, 2023). These and other events have demonstrated their capacity to shape social perceptions and behaviours. Their influence extends even to the ways in which young people interact with each other, as well as within their family and social circles (Baha, 2022).

These tools can contribute significantly to social good by facilitating community building, empowering people, promoting public interest messages, driving economic growth, and promoting justice and responsible consumption, among many other issues (Chang & Zhang, 2024). However, mass access to them has also served as a vehicle for the proliferation of misinformation and hate speech,

including xenophobic messages (Forrest & Wexler, 2023). In this context, the social network X and other similar platforms have played a central role, due to their viral design and the high volume of daily posts they host.

The structure and functioning of X provide an environment where xenophobic messages can quickly amplify. The brevity of posts, combined with algorithms that prioritise interaction and controversial content, favour the circulation of messages that appeal to intense emotions, such as fear or anger, which are inherent characteristics of hate speech. In addition, the partial anonymity offered by these platforms lowers the social barriers to discriminatory commentary, encouraging the proliferation of content hostile to minority groups, including migrants and refugees.

Recent studies on the social network X have highlighted a correlation between peaks of xenophobic activity on the platform and relevant social or political events in the real world, often linked to immigration or issues involving foreign nationals (Santana dos Santos et al., 2022). Within this framework, social networks function as both a barometer and a catalyst for xenophobic sentiments, particularly during significant political or global events. For instance, research has indicated that approximately 50% of tweets analysed during the COVID-19 pandemic constituted xenophobic harassment, with a substantial proportion directed at individuals perceived to be of Asian descent (Dhungana Sainju et al., 2022). Similarly, recent work such as that by Umarova et al. (2024) establishes comparable connections, based on an analysis of over 7,000 tweets related to US immigration policies and linked to accounts of influential figures in the public sphere.

X presents an ecosystem conducive to the proliferation of xenophobia. Nasuto and Rowe (2024) analysed 220,870 xenophobic messages on X concerning the discourse surrounding immigration in the UK, revealing a high degree of polarisation on this social network. Anti-immigration sentiment is notably more active and disseminates 1.66 times faster than pro-immigration messages. This indicates a significant presence of xenophobia in discussions pertaining to immigration-related events.

During periods of events related to immigration, the volume of publications demonising foreign groups increases substantially, frequently accompanied by misinformation and conspiracy theories. This underscores the critical role of these platforms as spaces where social tensions are both reflected and amplified.

Research has analysed the relationship between the design of digital platforms and the amplification of hate speech. The architecture of these platforms facilitates the rapid dissemination of hateful content, creating echo chambers that reinforce extremist views (Weber et al., 2023). This occurs because recommendation algorithms, designed to maximise user time and interaction, tend to favour polarising content, given its capacity to generate greater engagement. This algorithmic logic prioritises exposure to contentious messages, including those imbued with xenophobia. This is evident in digital platforms such as X, Instagram, and YouTube, where algorithms prioritise interaction over content moderation (Dutta, 2024).

Furthermore, mechanisms such as retweets, shares, and quick replies enable the instantaneous dissemination of offensive messages, reinforcing the perception that discriminatory views are more prevalent or accepted than is actually the case.

However, the perceived virulence or polarisation of messages on social networks is mitigated when the population has opportunities for real-life interaction with immigrant groups. This is supported by studies such as those by Arcila et al. (2022a), who compared the level of hate speech on social networks with the varying presence of affected groups in the locations from which the messages originated, using geolocation tools. Their findings indicate that in countries with a greater presence of immigrants, the volume of xenophobic messages on social networks such as X decreases. This suggests that physical interactions within the immediate environment play a significant role in shaping social perceptions of a group (Arcila et al., 2022a).

3. Objectives

The primary objective of this research is to analyse xenophobic messages disseminated on the social network platform X, covering the period from June to November 2024.

The following specific objectives were set in order to explore this issue in greater depth:

- To ascertain the degree of animosity within xenophobic messages disseminated on the social network platform X between June and November 2024.

- To characterise the textual content of messages negatively portraying foreigners or immigrants disseminated on the social network platform X between June and November 2024.
- To ascertain the impact and reach of hate messages disseminated on the X social media platform between June and November 2024, as measured by the number of likes and retweets received.

4. Methodology

This research adopts a mixed-methods (qualitative-quantitative) design to analyse and classify hate speech on social networks, with specific attention to the X platform within the Spanish context. This research is descriptive and exploratory, seeking to delineate the characteristics of a particular phenomenon, in this case, xenophobia and hate messages, and to explore trends and patterns in the data collected in 2024. It investigates patterns and describes the attributes of hate speech through statistical and content analysis.

Descriptive statistics (frequencies and percentages) were used to analyse the data. We analysed the discourses of 5 media outlets in Platform X: *ABC, El Mundo, El País, La Vanguardia,* and *20 Minutes*, from June to November 2024.

This study draws upon a dataset acquired from the Hatemedia Project (2021-2024), which employs a hate monitor to classify messages according to five levels of hate speech (ranging from 0 to 5). This monitor utilises an advanced data mining approach, analysing substantial volumes of social media posts through machine learning algorithms. The project methodology is structured in several phases:

- Data Collection: Messages posted on social networks in association with news media in Spain are collected. This encompasses mentions and comments pertaining to news items.
- Automatic Classification: Messages are processed via an algorithm that assigns a hate-level label based on the language and context of the message.
- Human Validation: A team of experts reviews and validates a sample of the classified data, ensuring the accuracy of the algorithm.
- Exploratory Analysis: Patterns and trends within the data are examined, such as the monthly distribution of messages or their relationship to media events.

The monitor also uses a library of hate speech, which includes up to five categories: general, political, sexual, misogynistic, and xenophobic. This allows different forms of hate speech to be identified and categorised with a high degree of accuracy. The work carried out by the hate monitor algorithm includes:

- Monthly frequency of hate messages, distributed by category.
- Classification of insults and derogatory terms.
- Analysis by referenced media.
- Distribution and retweets according to levels of hate.

Of particular interest is the determination of hate levels by the algorithm trained within the Hatemedia Project. To this end, the *Manual for Labelling Hate Messages* (De Lucas et al., 2022) was developed, which classifies these discourses into levels of intensity from 0 to 5, with the aim of analysing and categorising their impact and severity. This methodology ensures an objective assessment, grounded in linguistic and contextual criteria, enabling the identification of the potential consequences of the discourse analysed.

4.1. Levels of Hatred

The content of each of the five hate levels employed for the labelling of messages analysed by the hate monitor is detailed below:

Level 0: No Hate

Level 0 refers to messages that do not contain hate speech per se, although they might indirectly contribute to the generation of hate speech. These messages often use objective or neutral categories to describe a social group. For example:

> "A group of Moroccans have arrived at the border." While the mention of nationality is not inherently derogatory, its prominence as a primary characteristic could, within a biased or sensationalist context, contribute to prejudice.

Level 1: Indirect Stigmatisation

At this level, messages do not contain insults or overt verbal violence, but present facts or information that may reinforce stigmas towards a social group. These discourses are often related to the generalisation of negative behaviours. For example:

> "Immigrants take advantage of social benefits." Although there is no direct attack, the statement negatively generalises the intention of a group, promoting an unfavourable perception.

Level 2: Manifest Contempt

Messages classified at level 2 more directly attribute negative intentions to a group, often using abusive expressions or prejudices. At this level there are no explicit insults, but a discourse of rejection of the group is constructed. For example:

> "Muslims only come to impose their religion." This type of messaging conveys contempt for members of a group, using statements that seek to discredit them or minimise their legitimacy.

Level 3: Verbal Violence

Level 3 covers expressions in which verbal violence is evident, with insults, humiliation or contempt directed at one or more people because they belong to a social group. This level also includes linguistic intensifiers, such as "piece of shit" or "piece of", which increase the negative charge of the message. Example:

> "Morocco only exports shitty Moroccans who don't want to work." Here, the use of pejorative adjectives and derogatory generalisations clearly exemplifies hatred towards a group.

Level 4: Veiled Threats and Non-Direct Incitement of Violence

Level 4 messages contain indirect, implied, or veiled threats directed towards a social group or its members. These expressions may encompass intimidation, overt satisfaction regarding past harm, or indirect incitement to violence. While explicit incitement is not always present, the message aims to cultivate fear, humiliation, or a hostile environment towards the target group. Furthermore, expressions manifesting a desire to harm the group may appear, without directly implicating the sender in the desired actions. For example:

> "Be careful when you walk down the street. Not all fascists come home in one piece." This message uses an implicit threat that suggests a physical risk to the recipient, without direct incitement.

Another example would be:

> "I'm glad that bullfighter is dead. Bullfighting should disappear along with them." Here, although there is no direct threat or incitement, the celebration of death reinforces the atmosphere of hatred towards a social group linked to a cultural activity.

Level 5: Direct Incitement of Violence

Level 5 represents the most extreme point of hate speech and is characterised by explicit calls for physical violence, clear and direct threats, or explicit wishes for an individual or social group to suffer serious harm. These expressions may include references to actions such as killing, beating, raping, or physically assaulting members of a collective, or attacking institutions, symbols, or property related to them. For example:

"We should kill all those Muslims. They are destroying our country."

Another example would be:

"All the mosques must be burnt down. They are nothing more than nests of terrorists."

These messages not only openly incite violence, but also legitimise aggressive actions against the targeted groups, constituting a real danger of escalating into real-life violence.

Additional Examples

The methodology also contemplates complex situations, such as the use of humour or sarcasm, which could change the perception of hatred. For example:

"Those *Panchitos* look like Gremlins. If they get wet, they reproduce." In this case, although the humour is evident, the message continues to reinforce xenophobic stereotypes.

The tiered classification outlined above enables differentiation between messages that directly promote hate, insinuate it, or indirectly contribute to its proliferation. This approach provides a clear, systematic framework for analysing hate speech on digital platforms, ensuring consistent and reproducible labelling. Within this study, messages extracted from the social network X between June and September 2024 range from levels 0 to 3. No messages were found in categories 4 and 5, which represent overt incitement to violence.

5. Discussion of Results

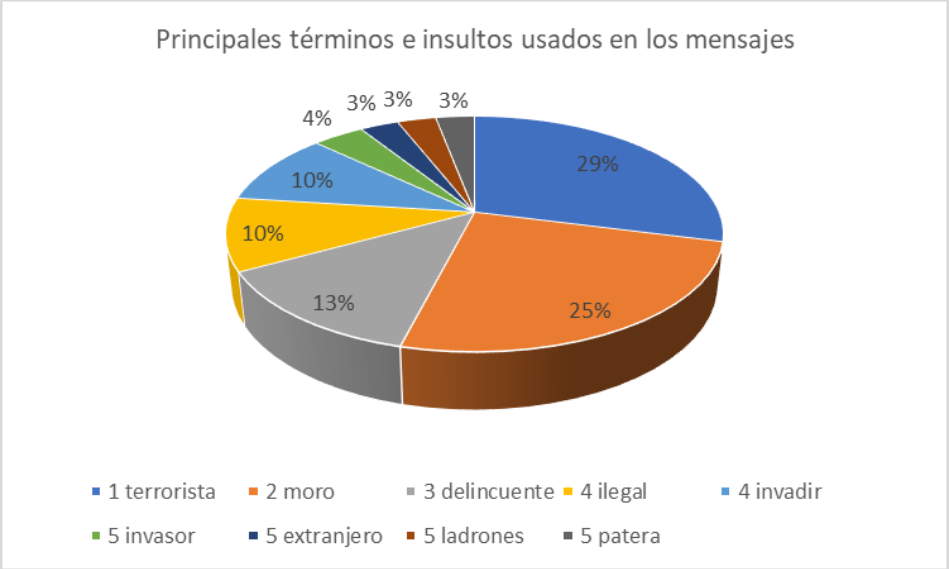
The results presented herein address the level of hatefulness in xenophobic messages, characterise negative representations of foreign or immigrant groups, and evaluate the impact of these messages based on their reception and propagation (likes and retweets). Table 1 and Graph 1 display the most frequent terms identified in xenophobic messages, along with their number of mentions and interactions (retweets and replies).

Table 1: Main terms and insults used in messages

Ranking	Term	Frequency	Percentage
1	Terrorist	9	29%
2	Moro	8	25%
3	Offender	4	13%
4	Illegal	3	10%
4	Invade	3	10%
5	Invader	1	4%
5	Foreigner	1	3%
5	Thieves	1	3%
5	Patera	1	3%

Source: Authors elaboration, 2024.

Figure 1. Main terms and insults used in messages



Source: Authors elaboration, 2024.

Regarding the most frequent terms identified in xenophobic messages, and considering the number of mentions, interactions, and percentage frequencies (retweets and replies), Graph 1 and Table 1 indicate a predominance of terms such as "terrorist" and "Moor." These constitute 29% and 25% of the total, respectively, which suggests a notable trend towards stigmatisation and negative categorisation within the examined discourse.

Conversely, terms such as "criminal" and "illegal/invader", representing 13% and 10% respectively, while less prevalent, remain significant and contribute to a narrative of criminalisation and dehumanisation. These findings emphasise how language employed in social media can influence and reinforce public policies and perceptions of minority groups, thereby promoting a biased and often negative view.

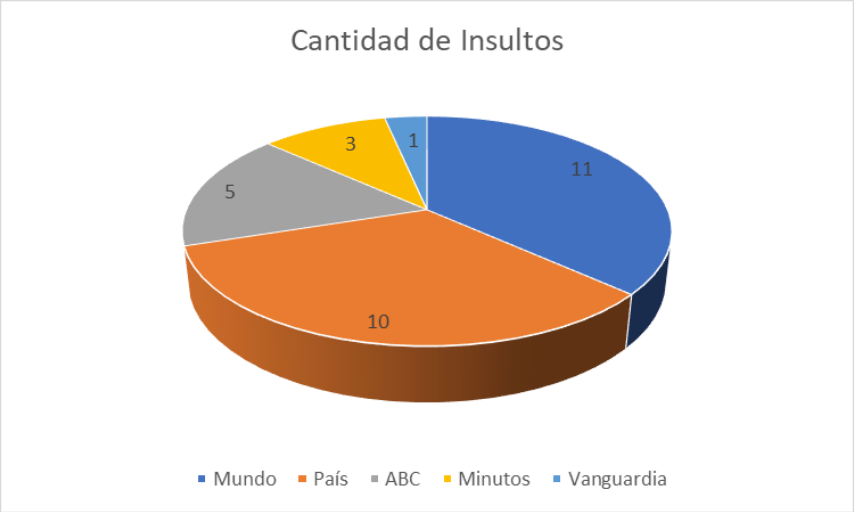
The analysis of xenophobic messages posted on the social network X between April and November 2024 indicates that certain derogatory terms, such as "terrorist", "Moor", and "delinquent", are recurrent and form part of a xenophobic narrative directed towards immigrant and foreign groups. Furthermore, while the impact of these messages, as measured through interactions such as retweets and mentions, is limited, the sheer volume of publications reinforces their visibility.

This aligns with findings from prior studies on misinformation and stigmatising messages regarding immigrants online (Sánchez Esparza et al., 2023), which identified recurring themes. Primarily, these themes concern the association of immigration with the economic burden imposed on the host country, where immigrants are perceived to be depriving nationals of social welfare and employment opportunities, the perceived threat to security attributing insecurity in various domains, including terrorism and crime, to the migrant population, and the purported risk to the preservation of the host culture's identity.

Regarding media sources, *El Mundo* and *El País* were identified as the most frequently referenced in messages containing insults. Continued monitoring of these messages is important to better understand the dynamics of hate speech on social networks and its relationship to the media agenda and socio-political events.

Graph 2 illustrates the total number of insults identified in the analysed messages, categorised by media outlet. *El Mundo* and *El País* exhibit the highest incidence of insults, followed by *ABC*.

Graph 2. Insults distributed by media outlets



Source: Authors elaboration, 2024.

In accordance with the objectives of the study to analyse xenophobic messages and to determine the level of hatred in these messages, it is observed that media outlets *El Mundo* and *El País* present the highest figures, with 11 and 10 insults respectively. This could indicate a higher frequency or intensity of xenophobic discourse transmitted through these media compared to others such as *ABC* (5 insults), *20 Minutes* (3 insults), and *La Vanguardia* (1 insult).

Table 2. Messages with the highest level of hate (Level 3)

Message text	Level of hatred	Related media
@abc_es @mikelayestaran In Moorland there will never be peace. That's the way it is.	3	ABC - Nov 24
@elmundoes I hope they kick out of Spain all those cowardly Ukrainians that we have put up with.	3	El Mundo - Nov 24
@elmundoes More cry babies. If they don't like where they live, they have it easy. Sell up and move on.	3	El Mundo - Nov 24
@elmundoes I don't give a shit about narco-boats. I have more problems with fence jumpers.	3	El Mundo - Nov 24
@elmundoes That's why they bring in so much "manpower". To do what?	3	El Mundo - Nov 24
@elmundoes Let them kill each other.	3	El Mundo - Nov 24
@elmundoes They abandon us, they steal, and we also have to put up with them in X. 👿	3	El Mundo - Nov 24
@elmundoes Ungrateful. They deserve to burn in the name of our European values.	3	El Mundo - Nov 24
@el_pais Libyan stuff...	3	El País - Nov 24
@20m Fuck the anti-Spanish pig who wants an invasion and the extinction of our people.	3	20 Minutes - Nov 24
@LaVanguardia Clear out El Raval of all anti-Westerners.	3	La Vanguardia - Oct 24
@elmundoes He has the face of a disgusting traitor. I hope they put him in the slammer, but where the Moors are.	3	El Mundo - Oct 24
@elmundoes We need mass deportations, withdraw nationalities given to repeat offenders, close the borders tightly.	3	El Mundo - Oct 24

@elmundoes Jews are you, you usurious scumbags.	3	El Mundo - Oct 24
@elmundoes The Jew had to be a bastard.	3	El Mundo - Oct 24
@elmundoes They shouldn't have come, they weren't fleeing from a war, they weren't fleeing from a war, they weren't fleeing from a war, they weren't fleeing from a war, they weren't fleeing from a war.	3	El Mundo - Oct 24
@20m To their fucking house, all of them.	3	El Mundo - Sep 24
@abc_es Wow... and I thought they were going to be Norwegians and it turns out they are Moroccans. I can't get over my astonishment.	3	ABC - Sep 24
@abc_es They bring us the bugs on purpose.	3	ABC – Aug 24
@20m Immigrants or narcos?	3	20 Minutes - Aug 24
@el_pais Send them back to their country and cut the crap.	3	El País - Aug 24
@el_pais Send them back to Morocco, there is no war there and they have their families.	3	El País - Aug 24
@el_pais INVADERS. It says INVADERS. INVADERS. INVADERS	3	El País - Aug 24
@elmundoes The nationalities of the criminals? No need, we already know.	3	El Mundo - Aug 24
@elmundoes In other words: 550 crimes a day go unpunished in Madrid. Establishment solution: import criminals from the recently emptied North African prisons...	3	El Mundo - Aug 24
@elmundoes We are already Africa https://t.co/4uSs48Rxp9.	3	El Mundo - Aug 24
@20m Let them take advantage if they serve to CONTROL the rubbish that comes illegally to Spain.	3	20 Minutes - Jul 24
@20m My, how the rubbish trucks have changed.	3	20 Minutes - Jul 24
@elmundoes Spain's image in the world of Moors and more Moors. Hahaha.	3	El Mundo - Jul 24
@elmundoes That doesn't stop them getting away with bullets and killing people who are having a quiet dinner. THIS SHOULD NOT HAPPEN. DEPORTATION NOW.	3	El Mundo - Jul 24
@elmundoes Well, if it has been stolen from a sub-Saharan, I guess it will be at the bottom of the sea by now.	3	El Mundo - Jul 24
@elmundoes You have turned France into a rubbish dump, no matter how much money you have, nothing but shit comes from Africa.	3	El Mundo - Jun 14
@elmundoes Algerians, Cameroonians, Moroccans, Senegalese, Malians... France's invaders want the Africanisation of Europe to go ahead.	3	El Mundo - Jun 14
@elmundoes Europe is a dunghill, first it should be cleaned up and then we'll see.	3	El Mundo - Jun 14
@elmundoes There's a lot of riffraff in Spain.	3	El Mundo - Jun 14

Source: Authors elaboration, 2024.

As shown in Table 2, the analysis of the messages published on the X platform classified as level 3 hate by the Hatemedia Project, shows a pattern of explicit hostility in the online communication, especially concentrated in certain media. In particular, *El Mundo* stands out for the repeated appearance of aggressively xenophobic and discriminatory messages, which could point to an audience that actively responds to this type of rhetoric.

For example, several messages in *El Mundo* in November 2024, such as "In Moorland there will never be peace. It is what it is." and "I hope they kick out of Spain all those cowardly Ukrainians that we have put up with", not only dehumanise the groups, but also promote a narrative of exclusion and hostility.

These messages share an overt contempt for immigrants and minorities, using pejorative terms and language that incites rejection and verbal violence.

Comparatively, other media such as *El País* and *La Vanguardia* also present hate messages, but in a significantly smaller quantity. An example from *El País* would be "Libyan stuff...", which, although less explicit than the messages in *El Mundo*, still contributes to negative stereotypes and perceptions. This suggests that, although hate speech is present in several media, the intensity and frequency varies considerably.

These differences in the quantity and nature of messages may reflect variations in editorial policies or comment moderation strategies of individual media outlets. This pattern underlines the importance of media outlets implementing more effective moderation practices and developing clear policies against the dissemination of hate speech, in order to mitigate the negative impact of hate speech on social cohesion and respect for diversity.

Table 3 shows the total number of messages and retweets classified according to hate levels (0 to 3), where level 0 represents barely perceptible speech and level 3 includes explicit hate speech.

Table 3. Messages and retweets by level of hatred

Level of hatred	Number of messages	Number of retweets
0	2906	158
1	395	6
2	251	13
3	955	38

Source: Authors elaboration, 2024.

Table 3 presents the distribution of messages and their retweets, categorised by the level of hate, from June to November 2024, on the social network X. The subsequent breakdown by level provides a lucid depiction of the dynamics of user interaction with hate speech on the platform.

Regarding level 0 (no hate), which encompasses neutral or non-offensive messages, it exhibits the highest number of messages (2906) and a relatively high number of retweets (158). This indicates that, despite the prevalence of these messages, their capacity to generate a reaction or virality is moderate, suggesting that users are more inclined to share content devoid of explicit hate.

Level 1 (indirect stigmatisation) messages are less frequent, with only 395 messages, and the corresponding number of retweets is exceptionally low (6). This may be interpreted as a reduced propensity for users to engage with content that, while not overtly aggressive, may contain negative insinuations or generalisations.

Concerning level 2 (overt contempt), with 251 messages, it also elicits a limited response in terms of retweets (13), which may reflect a reluctance of the audience to disseminate messages expressing contempt or rejection towards groups or individuals, notwithstanding the presence of a more explicit negative discourse.

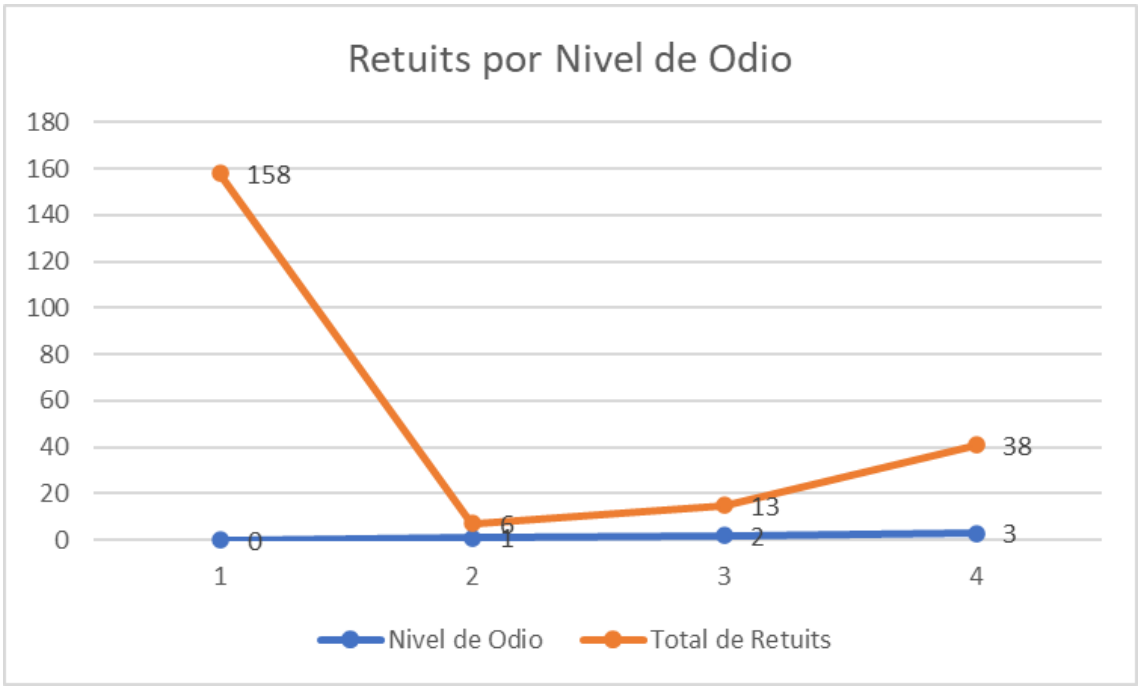
Level 3 (verbal violence) represents the second highest level in terms of the number of messages (955), but with only 38 retweets. This suggests that, although these explicit hate messages are numerous, the community does not frequently amplify them. This constitutes an encouraging indication that the most extremely negative content is not widely accepted or promoted by users.

The distribution of messages and retweets, as categorised by hate levels, suggests a potential awareness or rejection within the user community of more explicit and aggressive hate speech. Nevertheless, the substantial number of hate messages, particularly at levels 2 and 3, underscores the necessity for effective moderation and digital literacy strategies to mitigate the presence and impact of such content on social media platforms.

Figure 3 provides a comparison of the retweets generated by messages at hate speech levels 0, 1, 2, and 3. Level 0 messages, despite being less explicit, generate a greater number of retweets, whereas level 3 messages, which are more explicit, elicit less interaction. In this context, it is observed that neutral messages (Level 0) garner the highest number of retweets, potentially indicating a preference

for sharing content devoid of explicit hate. Conversely, messages with a more pronounced degree of hate (Level 3) are also shared, albeit to a lesser extent.

Figure 3. Retweets by level of hate



Source: Authors elaboration, 2024.

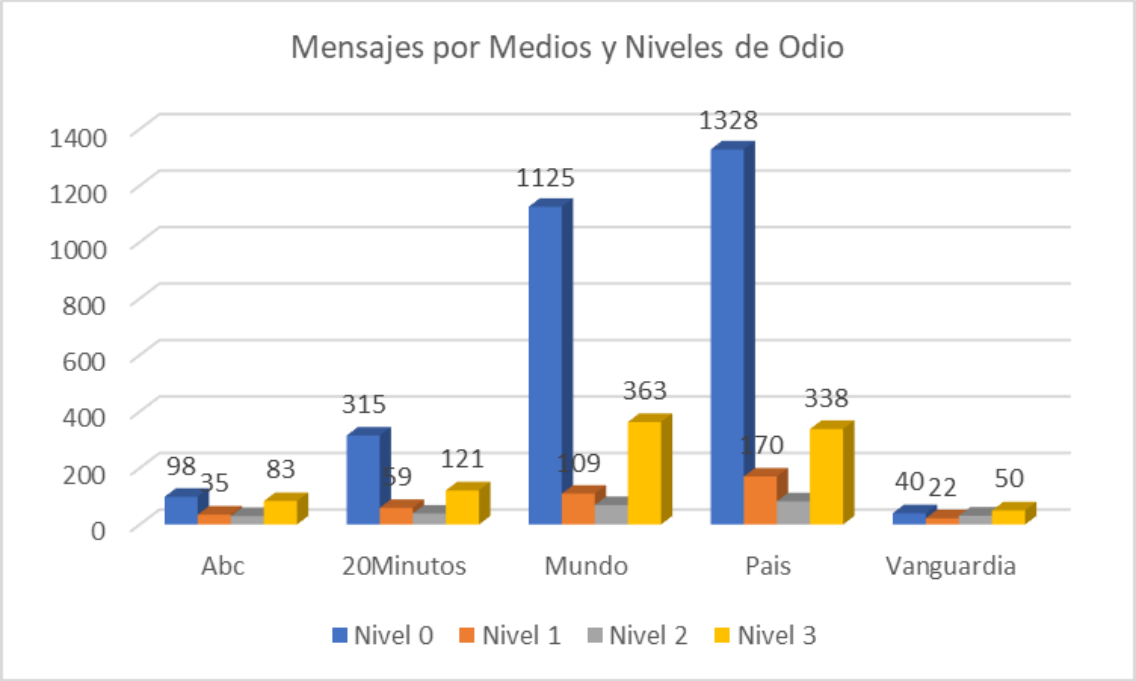
Table 4 and Graph 4 combines the levels of hate (0 to 3) and media, showing the number of messages classified at each level and for each media.

Table 4. Messages by media and levels of hatred

Medium	Level 0	Level 1	Level 2	Level 3
ABC	98	35	30	83
20 Minutes	315	59	39	121
El Mundo	1125	109	69	363
El País	1328	170	82	338
La Vanguardia	40	22	31	50

Source: Authors elaboration, 2024.

Figure 4. Messages by media and levels of hatred



Source: Authors elaboration, 2024.

The analysis shows that level 0, although less explicit, generates the most messages and retweets, suggesting that subtle hate speech is more likely to be shared. Higher levels, such as level 3, have fewer retweets, possibly due to their explicit nature limiting acceptance or visibility.

Media outlets such as *El País* and *El Mundo* have the highest volume of xenophobic messages. The discourses are amplified through mentions of news items published by these media and concentrate the largest number of messages at all levels, highlighting their influence on the media narrative that is replicated on social networks.

Regarding the monthly frequency of xenophobic messages, the distribution graph shows a significant increase in the months of June and September 2024. These peaks could be related to public events or news of general interest related to immigration issues. In this sense, some events have been in the months with the highest volume of xenophobic messages:

1. June 2024:
   - New Asylum Law: In June, Spain passed a new asylum law that made it easier for asylum seekers to obtain work and housing permits. This law generated widespread debate in the media and on social media.
   - Increase in arrivals by sea: In June, there was a significant increase in the number of migrants arriving on Spanish shores by sea, leading to concerns and discussions about the country's reception capacity.
2. September 2024:
   - Migration crisis in the Canary Islands: In September, the Canary Islands experienced a surge in migrant arrivals, leading to a humanitarian crisis and a debate on the management of immigration in Spain.
   - Immigration as the main concern: According to the September barometer of the Centre for Sociological Research (CIS), immigration became the main concern for Spaniards, surpassing unemployment. This change in public perception was also reflected in an increase in xenophobic messages on social networks.

The findings derived from X between June and November 2024 distinctly demonstrate the prevalence of hateful and xenophobic terms, such as "terrorist" and "Moor," which dominate the

discourse with high percentages of mentions and interactions. This predominance of specific terms underscores a trend towards the stigmatisation of certain ethnic and religious groups, a dynamic recognised by Fuentes Osorio (2024) and the United Nations (2019) as indicative of a society that facilitates discrimination through its language in the media and public platforms. The frequency of these terms not only reflects entrenched perceptions but also, according to the Council of Europe (1997), fosters attitudes of racism and xenophobia, thereby reinforcing invisible barriers against integration and mutual respect.

Furthermore, analysis of the interaction of these messages, specifically retweets and replies, indicates that, although the volume of xenophobic posts is significant, the overall interaction is relatively low, which could suggest a tacit resistance or disapproval on the part of the majority audience. However, the fact that these messages persist and are visible on media platforms such as *El Mundo* and *El País* reflects the urgent need for more rigorous moderation policies within the media, as suggested by research on the social impact of hate speech (Valle de Frutos, 2024).

The role of mainstream media in disseminating xenophobic messages and shaping public opinion is of considerable significance and should not be underestimated. The relationship between media narratives and the amplification of xenophobic discourse, as highlighted in the analysed messages, illustrates how media outlets can function as amplifiers of pre-existing prejudices, thereby fostering an environment in which hatred is not only perpetuated but also normalised (Hyewook, 2023). This phenomenon is particularly critical given the influence of artificial intelligence and machine learning in the detection and management of hate speech, as demonstrated in studies by Arcila-Calderón et al. (2022) and Gautam et al. (2024).

Continuous monitoring and analysis of these interactions are essential to better understand the dynamics of hate speech on social media and its relationship to socio-political and media events. Digital education and effective intervention strategies must be prioritised to combat the proliferation of hate speech, thus ensuring a more inclusive and respectful platform. This approach not only responds to the immediate need to mitigate the impact of online hate but also paves the way for deeper reflection and long-term solutions that address the cultural and social roots of the problem.

6. Conclusions

The research on hate speech on the social network platform X has successfully achieved its stated objectives, providing a comprehensive and detailed insight into the nature and impact of these messages. Initially, in determining the level of hate within xenophobic messages, it was identified that highly negative terms, such as "terrorist" and "Moor," predominated in the communication, representing 29% and 25% of the total mentions, respectively. These findings indicate a considerable level of hostility directed towards specific groups, reflecting an environment in which prejudice and stigmatisation are prominent.

With regard to the characterisation of the texts, the research revealed that the language employed in the messages not only categorises immigrants and foreigners pejoratively but also promotes a narrative of criminalisation and dehumanisation. Terms such as "criminal," "illegal," and "invading," while less frequent than more overtly aggressive terms, contribute significantly to a distorted and negative perception of these groups. This characterisation of hate messages underscores the urgency of addressing the analysis of how language in digital media can influence social attitudes and perpetuate discrimination.

The findings demonstrate how clichés concerning immigrants are reproduced and recurrently appear in fake news, which are intentionally and coordinately disseminated on social networks within host countries, such as Italy, Greece, and Spain (Sánchez Esparza et al., 2024). Consequently, the image of the immigrant becomes associated with attributes such as vandal, anti-social, usurper, or criminal, leading to the construction of systems of meaning (myths) and social imaginaries that stigmatise this group and portray them as an unwelcome element within the host society.

When assessing the impact and repercussion of these messages, it is observed that, despite the high volume of xenophobic posts, interaction in terms of likes and retweets is relatively low. This could be interpreted as a reluctance on the part of most users to increase levels of hate speech, although the continued presence of these messages remains a concern. It is notable that, although messages with high hate content such as those classified in level 3 are frequent, their rate of retweets is not proportionally

high, suggesting a possible dissonance between the visibility of the content and the online community's approval of it.

The results also imply that mainstream media outlets, such as *El Mundo* and *El País*, play a significant role in the dissemination of these discourses, raising questions about editorial responsibility and content moderation policies on powerful platforms. The association between the intensity of hate speech and these media indicates the need for strategies to limit the spread of xenophobic messages, which not only affect the individuals directly, but also poison the wider social environment.

This study underlines the importance of continuing to monitor and analyse hate speech on social media to better understand its mechanisms and effects on society. Looking ahead, it is essential to implement and strengthen educational, moderation, and public policy measures that can mitigate the harmful effects of online hate, promoting a more inclusive and respectful digital environment. This will not only reduce the prevalence of such harmful speech but also foster a culture of respect and tolerance in the global digital sphere.

Importantly, the study's focus on messages collected specifically from the social media platform X limits the generalisability of the results to other social networks, where patterns of hate speech may vary due to differences in user demographics, moderation policies, and platform design. Furthermore, while the use of machine learning algorithms to classify hate speech provides a valuable tool, such algorithms may lack the capacity to fully comprehend context and tone, particularly in instances of sarcasm or irony, potentially resulting in misclassifications.

In conclusion, this study on hate speech found on the social network platform X demonstrates the complexity and the imperative to address this phenomenon within our digital societies. Despite methodological and scope limitations, the results obtained underscore the necessity for multidisciplinary interventions, including technological improvements, more rigorous public policies, and effective educational strategies to combat hate speech. The persistence of pejorative terms and xenophobic messages, and their impact on social cohesion, challenge us to reflect on our collective responsibility in creating more inclusive and respectful digital spaces. This endeavour requires not only constant vigilance of platforms and media but also the active engagement of every user to foster constructive and respectful dialogue, ensuring that technology serves as a bridge, rather than a barrier, to mutual understanding and human dignity.

7. Acknowledgements

This paper is a partial result of the project “Hatemedia: Taxonomy, Presence and Intensity of Hate Speech in Digital Environments Linked to Spanish Professional Media.” Hatemedia (PID2020-114584GB-I00), funded by MCIN/AEI/ 10.13039/5011000110.

References

Aldamen, Y. (2023). Xenophobia and Hate Speech towards Refugees on Social Media: Reinforcing Causes, Negative Effects, Defense and Response Mechanisms against That Speech. *Societies, 13(*4), 83. https://doi.org/10.3390/soc13040083

Arcila-Calderón, C., Sánchez-Holgado, P., Quintana-Moreno, C., Amores, J., & Blanco-Herrero, D. (2022). Hate speech and social acceptance of migrants in Europe: Analysis of tweets with geolocation. [Discurso de odio y aceptación social hacia migrantes en Europa: Análisis de tuits con geolocalización*]. Comunicar, 71*, 21-35. https://doi.org/10.3916/C71-2022-02

Arcila-Calderón, C., Amores, JJ, Sánchez-Holgado, P., Vrysis, L., Vryzas, N., y Oller Alonso, M. (2022). ¿Cómo detectar el odio en línea hacia migrantes y refugiados? Desarrollo y evaluación de un clasificador de discurso de odio racista y xenófobo mediante aprendizaje superficial y profundo. *Sustainability, 14* (20), 13094. https://doi.org/10.3390/su142013094

Baha, O. A. (2022). Nurturing Communication through Social Online Platforms. *International Journal of Humanities and Education Development (IJHED), 4(1),* 155-159. https://doi.org/10.22161/

Bruneel, S., De Wit, K., Verhoeven, J. C., & Elen, J. (2013). Facebook: When education meets privacy. *Interdisciplinary Journal of e-Skills and Lifelong Learning, 9,* 125-148. https://doi.org/10.28945/1868

Bursztyn, L., Egorov, G., Enikolopov, R., & Petrova, M. (2019). Social media and xenophobia: Evidence from Russia (Working Paper No. 26567). *National Bureau of Economic Research*. https://doi.org/10.3386/w26567

Bursztyn, L., Egorov, G., Enikolopov, R., & Petrova, M. (2020). Social Media and Xenophobia: Evidence from Russia. *CEPR Discussion Paper No. DP14877*, Available at SSRN: https://ssrn.com/abstract=3628198

Chang, L., & Zhang, X. (2024). Introduction: Platforms for social good. *Global Media and China, 9(3*), 253-259. https://doi.org/10.1177/20594364241241777

Comisión Europea contra el Racismo y la Intolerancia. *Recomendación de política general N.° 15 de la ECRI sobre la lucha contra el discurso de odio*; Consejo de Europa: Estrasburgo, Francia, 2016.

Consejo de Europa. *Recomendación nº R 20 del Comité de Ministros a los Estados miembros sobre el "discurso de odio"* ; Consejo de Europa: Estrasburgo, Francia, 1997.

De Lucas Vicente, A; Römer Pieretti, M.; Izquierdo, D.; Montero-Diaz, J.; Said-Hung, E. (2022). Manual para el Etiquetado de mensajes de odio. figshare. Presentation. https://doi.org/10.6084/m9.figshare.18316313.v4

Dhungana Sainju, K., Zaidi, H., Mishra, N., & Kuffour, A. (2022). Xenophobic Bullying and COVID-19: An Exploration Using Big Data and Qualitative Analysis. *International Journal of Environmental Research and Public Health, 19*(8), 4824. https://doi.org/10.3390/ijerph19084824

Dutta, M. J. (2024). Digital platforms, Hindutva, and disinformation: Communicative strategies and the Leicester violence. *Communication Monographs*, 1–29. https://doi.org/10.1080/03637751.2024.2339799

Forrest, K. B., & Wexler, J. (2023). Social media platforms. In K. B. Forrest & J. Wexler (Eds.), *Is justice real when reality is not?* (pp. 147-159). Academic Press. https://doi.org/10.1016/B978-0-323-95620-8.00001-3

Fuentes Osorio, J. L. (2024). Hateful speech. La expansión del discurso de odio. *Revista Electrónica de Criminología. 02-08*. 1-30.

Gautam, A., Singh, A., Verma, A., & Sinha, J. (2024). Hate speech detection using deep learning. *International Journal for Research in Applied Science & Engineering Technology (IJRASET)*, 12(5), 196–202. https://doi.org/10.22214/ijraset.2024.61475

Gelashvili, T. (2018). *Hate speech on social media: Implications of private regulation and governance gaps* (Master's thesis, Lund University). Faculty of Law, Lund University.

Hyewook J. (2023). Hate Speech, Subject Agency and Performativity of Bodies. *The Journal of Criticism and Theory, 28(1)*, 271-313. 10.19116/theory.2023.28.1.271

Kaur, T. (2023). Digital platforms: Social media platforms, knowledge platforms, media sharing platforms, service-oriented platforms. *International Journal for Research in Applied Science & Engineering Technology, 11(VII*), 2031. doi: 10.22214/ijraset.2023.54879

Martínez-Rolán, X., Sierra, J., & Ring Carlson, C. (2024). Discursos de odio, populismo digital y autoritarismos en red. *Revista Latina De Comunicación Social, (82).* Recuperado a partir de https://nuevaepoca.revistalatinacs.org/index.php/revista/article/view/2298

Ministerio del Interior de España (2020). Informe de Evolución de los Delitos de Odio en España. https://bit.ly/3PGN4ys

Ministerio de Inclusión, Seguridad Social y Migraciones. (2017). *Protocolo y sistema de indicadores para la detección del discurso de odio en redes sociales*. https://bit.ly/3QImKnr

Nasuto, A. & Rowe, F. (2024a). Exposing Hate - Understanding Anti-Immigration Sentiment Spreading on Twitter. *arXiv.org, abs/2401.06658* doi: 10.48550/arxiv.2401.06658

Nasuto A, Rowe F (2024b) Understanding anti-immigration sentiment spreading on Twitter. *PLoS ONE 19*(9): e0307917. https://doi.org/10.1371/journal.pone.0307917

Observatorio Español del Racismo y la Xenofobia (Oberaxe). (2017a). *Evolución del racismo, la xenofobia y otras formas de intolerancia en España.* Ministerio de Trabajo, Migraciones y Seguridad Social.

Observatorio Español del Racismo y la Xenofobia (Oberaxe), Centro de Investigaciones Sociológicas. (2017b). *Actitudes hacia la inmigración X*. Estudio Nº 3190. https://bit.ly/3ohwSqD

Observatorio Español del Racismo y la Xenofobia [OBERAXE]. (2024). *Informe anual de monitorización del discurso de odio en redes sociales 2023.* Ministerio de Inclusión, Seguridad Social y Migraciones. https://www.inclusion.gob.es/oberaxe/es/index.htm

Organización de las Naciones Unidas (ONU) (2019). *La estrategia y plan de acción de las Naciones Unidas para la lucha contra el discurso de odio*. https://bit.ly/3ibIWpU

Raborife, M., Ogbuokiri, B., Aruleba, K. (2024). The Role of Social Media in Xenophobic Attack in South Africa. (2024). *Journal of the Digital Humanities Association of Southern Africa* , *5*(1). https://doi.org/10.55492/dhasa.v5i1.5026

Sánchez Esparza, M., Diéguez, I. V., & Arribas, A. M. (2023). Mapping Stigmatizing Hoaxes Towards Immigrants on Twitter and Digital Media: Case Study in Spain, Greece, and Italy. In E. Said Hung & J. Diaz (Eds.), *News Media and Hate Speech Promotion in Mediterranean Countries* (pp. 136-161). IGI Global Scientific Publishing. https://doi.org/10.4018/978-1-6684-8427-2.ch008

Sánchez Esparza, M., Vázquez Diéguez, I., & Merino Arribas, D. (2024). La semiótica del odio en los bulos sobre inmigrantes detectados por plataformas de fact checking en España, Grecia e Italia. *Revista ICONO 14. Revista científica De Comunicación Y Tecnologías Emergentes, 22(1*), e2083. https://doi.org/10.7195/ri14.v22i2.2083

Santana dos Santos, A.R.; De Oliveira Rodrigues, C.M. & Summer de Melo, H.B. (2022). Identifying Xenophobia in Twitter Posts Using Support Vector Machine with TF/IDF Strategy. In *Proceedings of the XVIII Brazilian Symposium on Information Systems (SBSI '22). Association for Computing Machinery, New York, NY, USA, Article 37*, 1–7. https://doi.org/10.1145/3535511.3535548

Schäfer, S., Rebasso, I., Boyer, M. M., & Planitzer, A. M. (2024). Can We Counteract Hate? Effects of Online Hate Speech and Counter Speech on the Perception of Social Groups. *Communication Research, 51(5),* 553-579. https://doi.org/10.1177/00936502231201091

Umarova, K., Okorafor, O., Lu, P., Shan, S., Xu, A., Zhou, R., Otiono, J., Lyon, B., & Leshed, G. (2024). Xenophobia Meter: Defining and Measuring Online Sentiment toward Foreigners on Twitter. *Proceedings of the International AAAI Conference on Web and Social Media, 18*(1), 1517-1530. https://doi.org/10.1609/icwsm.v18i1.31406

Valle De Frutos, S. (2024). Discurso de odio y ofensivo en la red social Twitter hacia el colectivo chino. Análisis de la sinofobia: del rechazo cultural encubierto al explícito. *Anuario Electrónico de Estudios en Comunicación Social "Disertaciones", 17(1).* https://doi.org/10.12804/revistas

We Are Social. (2024). *Digital 2024*. We Are Social. https://wearesocial.com/es/blog/2024/01/digital-2024/

Weber, I., Vandebosch, H., Poels, K., & Pabian, S. (2023). Features for hate? Using the Delphi method to explore digital determinants for online hate perpetration and possibilities for intervention. *Cyberpsychology, Behavior, and Social Networking, 26*(7). https://doi.org/10.1089/cyber.2022.0195