Journalists' Ethical Responsibility: Tackling Hate Speech against Women Politicians in Social Media through Natural Language Processing Techniques

Maria Iranzo-Cabrera
Universitat de València, València, Spain
https://orcid.org/0000-0002-6237-6041
maria.iranzo-cabrera@uv.es

Maria Jose Castro-Bleda
Universitat Politècnica de València, València, Spain
https://orcid.org/0000-0003-1001-8258
mcastro@dsic.upv.es

Iris Simón-Astudillo
Universidad de Valladolid, Valladolid, Spain
https://orcid.org/0000-0003-3114-8414
iris.simon@uva.es

Lluís-F. Hurtado
Universitat Politècnica de València, València, Spain
https://orcid.org/0000-0002-1877-0455
lhurtado@upv.es

Abstract

Social media has led to a redefinition of the journalist's role. Specifically on Twitter, these professionals assume an influential position and their discourse is dominated by personal opinions. Taking into consideration that this platform has proven to be a breeding ground for polarization, digital harassment and hate speech, notably against women politicians, this research aims to analyze journalists' involvement in this complex scenario. The investigation aims to determine whether, immersed in online and gender defamation campaigns, journalists enhance the quality of public debate or, on the contrary, they reinforce the visibility of this hostile content. To this end, we examined a sample of 63,926 tweets published from 23 to 25 November 2022 related to a campaign of political violence against the Spanish Minister of Equality using Natural Language Processing tools and qualitative content analysis. Results show that during those three days, at least half of the tweets contained hate speech and improper language. In this climate of hostility, journalists participating in the debate not only have an ability to attract likes and retweets, but also exhibit polarization and use hate speech. Each ideological position -for and against the Minister- is also reflected in their own uncivil strategies. Under the umbrella of free speech and regardless of argumentative discourses, those journalists who lean towards ideological progressivism tend to insult their opponents, and those on the political right use divisive constructions, stereotyping and irony as attack techniques.

Keywords

hate speech; improper language; polarization; gender; Twitter; social media; journalism; political communication; Natural Language Processing

To cite this article, please refer to: Iranzo-Cabrera, M., Castro-Bleda, M. J., Simón-Astudillo, I., & Hurtado, L.-F. (2024). Journalists' Ethical Responsibility: Tackling Hate Speech Against Women Politicians in Social Media Through Natural Language Processing Techniques. *Social Science Computer Review, 0*(0). https://doi.org/10.1177/08944393241269417.

Introduction

Social media, especially Twitter (Molyneux & Mourão 2019), has become a daily tool in the work of journalists during the last two decades (Mills et al. 2021). For seven-in-ten U.S. journalists it is the social media site they use most or second most for their job (Jurkowitz & Gottfried 2022) and, in the case of Spain, 85% of journalists have an account in this social network (nowadays known as X) (Asociación de la Prensa de Madrid 2022, p. 70). The academic community has emphasized its functionalities in journalistic production process, as it allows for gauging societal interest, facilitates information collection, and enables the selection of alternative sources. In addition, it has brought innovation in the content dissemination phase and in the interaction with audiences. This technological platform enables bidirectional communication between journalists and the public in a disintermediated society (Orihuela 2015), where any user can create and disseminate their own contents autonomously without intermediaries, such as media (Iranzo-Cabrera & Casero-Ripollés 2023).

However, this technology has also resulted in a redefinition of the journalist's role. In this digital space, journalists distance

themselves from the media they work for and assume an influential position as a "facilitator, guide, and even mentor" to an audience that comprises all family, friends, and followers (Garcia & Marta-Lazo 2017, p. 92). This effort to connect with users through humanization, emotions and personalization places journalists in dilemmas such as being factual or expressing their opinions, as well as sharing professional or personal information (Brems et al. 2017).

This change in attitude towards the traditional journalistic ethics that demand objectivity and independence from professionals (Kovach & Rosenstiel 2007) occurs precisely in a context of a crisis of authority and credibility in journalism (Michailidou & Trenz 2021; Ward 2018, 2020). Not only that, but social media has proven to be a breeding ground for polarization and its expressiveness through digital harassment and hate speech (Carr & McCracken 2018). Relevant studies such as Piñeiro Otero & Martínez-Rolán (2021) have noted the aggressiveness of the attacks against feminism on Twitter, and others have highlighted that women with influence in the public sphere, particularly politicians and journalists, are among the groups that receive the most hate in this digital agora (Shane et al. 2022; Miškolci et al. 2020; Bhat 2024).

Considering these reasons, this research aims to analyze the role of journalists participating in this complex scenario of polarization and online gender violence on Twitter. It seeks to diagnose whether, immersed in defamation campaigns, they enhance the quality of public debate or, on the contrary, journalists reinforce the visibility of these hostile contents limiting their reporting to the repetition of hate speeches verbalized by others or even by themselves. Furthermore, this investigation wants to ascertain whether news pieces act as validation or legitimization for the public expressing hate speech on this network.

According to UNDP's Gender Social Norms Index (2023), nearly half the world's people believe that men make better political leaders than women do. Moreover, a survey answered by 60 members of the European Parliament or MEP staff showed that online political hate is highly gendered conducted (Steinert et al. 2023). Women parliamentarians are also more exposed to hateful comments about their private life (31% of women vs. 10% of men MEPs) or their physical appearance (41% of women vs. 29% of men MEPs). For that matter, aggressors may use the marginalized and stigmatic identities of Black and Asian women to boost their own profiles (Rodis 2021). In the case of African women parliamentarians, 46% received sexist attacks via the Internet or social media (2021). In this line, several political leaders have been subjected to hate campaigns on the Internet, viewed through a gender lens and not judged solely on their performance. This is the case of the former Prime Minister of Finland, Sanna Marin; the former Prime Minister of New Zealand, Jacinda Ardern; or the Brazilian federal deputy Manuela D'Ávila.

To carry out the study about the promotion of gender hate speech from a journalism perspective, we have chosen one of the hate campaigns generated on Twitter against a woman politician, the Spanish Minister for Equality from 2020 to 2023 and member of Podemos political party, Irene Montero. Podemos, together with PSOE and other left-wing formations, formed a left-wing coalition in January 2020 to take over the government of Spain. Since then, this politician has been the target of several hate campaigns on the Internet for being the partner of one of the founders of her party -Pablo Iglesias- and for being the visible head of the Ministry of Equality, thus promoting gender political changes (Durántez-Stolle et al. 2023; Martínez-Sanz et al. 2024; Arencón Beltrán et al. 2023). Among the legal changes encouraged by Montero, it is worth highlighting the Organic Law on the Integral Guarantee of Sexual Freedom –commonly known as the "Only Yes Means Yes" law–, which is based on the need for clear and unequivocal sexual consent. According to the new law, silence or lack of resistance from a victim could not be interpreted as an approval for intimacy. Furthermore, it eliminated the distinction between sexual abuse and sexual assault.

The controversy was caused because a new range of sentences was established and, in some cases, the minimum ones were lower, which has led to some convicted sexual offenders benefiting from downward sentence reviews. Right wing and far-right politicians have used legal changes such as this and others related to abortion rights and the recognition of transgender people to launch continuous bitterly sexist and personal attacks on her, her party and feminism (Jones 2022). One of the MPs most critical of these feminist ideas has been Carla Toscano, representative of the far-right party VOX. Opposed to abortion, euthanasia and laws against gender-based violence, "she tends to express her opinions in a very blunt manner, both in her writings and in her presentations' (Corbal 2022). Toscano often makes her opposition to feminism visible by wearing T-shirts with slogans such as '#Notmetoo', 'I love Patriarchy' or 'Stop feminazis'.

In this context of increasing political and social polarization around gender issues (Arencón Beltrán et al. 2023; Cabezas Fernández et al. 2023), one of the most significant scandals happened inside Spanish Parliament, a controversy that was immediately transferred to Twitter. On November 23, 2022, during her speech in the Lower House, the far-right deputy Carla Toscano described the Minister as a rapist's liberator and pointed to their private life: "the only merit you have is having deeply studied Pablo Iglesias" (Hermida 2022; Sánchez-Meza et al. 2023), alluding to a career advancement for being her partner.

For our purpose, the incorporation of Natural Language Processing (NLP) tools plays a pivotal role as a methodology for systematically identifying and analyzing hate speech and related issues like improper language within Twitter discourse. Leveraging NLP techniques and deep learning algorithms enables us to automate the detection process, efficiently sifting through vast amounts of textual data to pinpoint instances of hate speech. Furthermore, this automated approach will be compared and contrasted with manual analysis, conducted by human annotators, to validate the accuracy, efficiency, and

nuanced understanding achieved by the NLP-driven hate speech detection methodologies. This comparative analysis serves as a crucial step in not only validating the effectiveness of the automated NLP tools but also in understanding potential discrepancies and refining the algorithms to better align with human comprehension and interpretation of hate speech within the social media discourse. In addition, special emphasis is placed on detecting the presence of journalists and the media inside this digital sphere.

Twitter, Hate Speech, and Journalism

Polarization and Hate Speech

Under the banner of freedom of speech, social media have been considered a kind of online agora (Konikoff 2021) because it allows any citizen to express themselves. However, in recent years, it has become evident that these channels contribute to the dissemination of content that clashes with democratic principles, amplifies more radical viewpoints than face-to-face interactions (Colleoni et al. 2014) and facilitates the infiltration of hate speech (Llorca-Asensi et al. 2021; Acosta-Quiroz & Iglesias-Osores 2020; Alonso-González 2019; Rodríguez-Fernández 2019). Twitter is one of these social networks that has gone from being a platform for hope to an arena for polarization and hate (Blanco-Alfonso et al. 2022).

This communication phenomenon on social media, coupled with a climate of prejudice and intolerance permeating contemporary societies, leads to polarization (Fletcher et al. 2020). It compounds the alignment between political and social identities, reinforcing stereotypes and negative perceptions of outgroups (Wilson et al. 2020). Within this broad concept, we can distinguish ideological polarization, at the level of political views, and affective polarization, that refers to the intensification of positive attitudes towards one group or idea and negative attitudes towards another group or idea. The latter has "given rise to othering and distrust of the opposition, in addition to claims of moral superiority for one's own side" (Kreiss & McGregor 2024).

It explains why some content published within these 'social' spaces presents a message of inferiority or antipathy toward certain social groups (Miškolci et al. 2020), increases prejudices, consolidates resentment and plays a fundamental role in the escalation toward violence (Leader-Maynard & Benesch 2016). In fact, Vale & Serra (2019) have highlighted the increase in negative words and insults from political actors, as well as Niñoles Galvañ and Ortega-Giménez (2020) have demonstrated the normalization of hate speech in their messages. This is especially evident in two-party political systems with pluralistic electoral rules (Urman 2020; Cinelli et al. 2021; Hawdon et al. 2022) such as the Spanish one, especially after the emergence of far-right political groups that favor the dissemination of hate speech against certain social groups like migrants, the LGBTQ+ community, and feminists (Salvador 2021).

Particularly on Twitter, "a greater presence of strategies and dynamics of polarized discourse and segregation" (Peña-Fernández et al. 2023, p. 58), as well as "echo chamber behaviors"' (Flamino et al. 2023, p. 904), have been detected, mostly when exploring divisive hashtags (Bruns 2021). This does not preclude cross-ideological exposure on Twitter, though without a substantial contribution to the deliberative democracy (Matuszewski & Szabó 2019). Rather, such practices tend to drive the debate towards "a forum of attacks, obloquies and insults based on emotionality" (Hernández-Santaolalla & Sola-Morales 2019, p. 117). This tendency has increased since Elon Musk bought this social media in 2022. Since that time Twitter's engagement-based ranking algorithm amplifies emotionally charged content, above all angry tweets (Milli et al. 2023). It has also enabled suspended accounts, such as Donald Trump's, to be activated or to be recognised as an official account for a fee, as it did with QAnon, one of the main conspiracy theories of the US far right. Musk "has been vocal about being a 'free speech absolutist' who believes in unfettered discussions online" (Frenkel & Conger 2022).

At this point, we must focus on the lack of a unanimous definition of hate speech, due to its high degree of subjectivity, which can make detection more difficult (Hawdon et al. 2023). The perception of these hate constructs displays an ideological bias, thus, detection is selective and only occurs when formulated by a group with different ideological affiliations (Abuín-Vences et al. 2022). It is difficult to perceive because Twitter users tend to follow homophilic political actors (Barberá 2014) and people within their social characteristics (Zhang & Ho 2022). Furthermore, these discursive strategies of polarization can be gratifying; that is, hate speech involves a self-righteous feeling of moral superiority (Nikolaev et al. 2023), as well as a sense of community that can trigger political anger (Cheng et al. 2023).

In spite of this complex and subjective reality, recent studies consider an expression of hate as a conscious and voluntary public statement intended to denigrate a group of people or incite discriminatory, hostile or violent harm toward a particular group or a person belonging to it (Blanco-Alfonso et al. 2022; Paz et al. 2020; Gagliardone et al. 2015). This type of action can be expressed not only through rhetorical strategies, but also non-verbally or symbolically, as well as through metaphorical or ambiguous terms aimed at encouraging criminal actions or promoting prejudice and intolerance toward others (Lingiardi et al. 2020). These less explicit verbal exclusionary strategies include lack of argumentation, fallacies and narrative traps, the homogenization of certain groups, stereotyping, the inclusion or exclusion of specific terms, the us-versus-them dichotomy, approaches based on conflict, and humor (Martínez-Sanz et al. 2024; Noriega & Iribarren 2014; Tortajada et al. 2014). The construction of hate speech is based on a violation of civic and social norms; the induction of shame in the victims, as well as inducing it through threats and intimidation; an attempt to dehumanize the victims by comparing them to animals; and through

misinformation towards those people or groups to which they belong (Williams 2021).

Within the realm of hate circulating in the digital sphere, numerous studies have pointed to a highly gendered and sexualized type of hate, defined as online misogyny. It can be explained as the harassment of women on the Internet, mostly on Twitter, through abusive and sexist language or imagery, as well as threats of violence (Ging & Siapera 2019; Massanari 2017). Common expressions of online misogyny include body shaming, sexualization, objectification, the use of harmful stereotypes such as the weak nature of women (Jones 2016), infantilizing and patronizing language, rape threats, and dehumanization (Akbar & Safdar 2023). What exacerbates this situation is the significant harm of being able to mention the women targeted by these attacks on Twitter, making them participants in their own victimization.

While this online hate is most often perpetrated by someone unknown, this climate of criticism reaches a new level of visibility when it extends to institutional discourse and media coverage (Sánchez-Meza et al. 2023). In this regard, the so-called pseudo-media or far-right alternative media, websites that simulate media outlets and violate elemental journalistic ethics, have been characterized by the hatred injected against women and feminism (Palau-Sampio & Carratalá 2022). In the last decade, a significant number of pseudo-informational websites have emerged connected to far-right ideology (Figenschou & Ihlebaek 2019). Through mundane lies and vague or scant, repetitive and easy-to-understand content (Kim & Gil de Zúñiga 2021), these platforms appeal to the hatred of their audiences (typically men) towards vulnerable groups (women, immigrants, LGBTQ+ people) through the so-called "journalism of resentment" (Kimmel 2019). The lack of transparency in their ownership and authorship, their aggressive editorial line (Del-Fresno-García 2019) and their commitment to presenting value judgements as facts evidence "the fraudulent nature of the proposals that are presented as alternatives to conventional media" (Palau-Sampio & Carratalá 2022, p. 3).

Taking these arguments into account, we then ask the following:

RQ1: What proportion of hate and offensive language is detected in the tweets generated around Irene Montero?

RQ2: Are journalists and media accounts participating in this hate campaign on Twitter? Do they make use of hate speech strategies?

Freedom of Expression and Journalistic Ethics

In terms of laws, the European Union has urged its member states to protect citizens from hate speech. In Spain, the Penal Code, in its Title XXI, criminalizes and penalizes various conducts related to incitement to hatred and discrimination, especially against groups or individuals on grounds of racial or ethnic origin, religion, sexual orientation, gender identity, among others. However, when it comes to journalism, the concept of hate speech, including media violence, is not yet recognized in its self-regulation codes (Sánchez-Meza et al. 2023). In response, journalists' associations argue that the ethical principles of journalism are incompatible with hate speech (Asociación de la Prensa de Madrid 2024).

From a libertarian perspective, avoiding the publication of certain opinions –including those with hate speech– is seen as self-censorship, and therefore a threat to democracy (Cohen-Almagor 2013; George 2014; Slagle 2009). Those who hold this view contextualize their argument in an era of growing polarization in which the term "hate speech" is considered a catch-all rhetorical weapon against opponents, and journalists are expected to allege moral equivalency between both sides (Johnson et al. 2021). In contrast, media ethics hold that journalists have a social obligation to establish ethical standards that enable acceptable public discourse (Abuín-Vences et al. 2022). Only in this way the necessary conditions for deliberative democracy based on freedom of expression, truth, integrity, and tolerance can be created (Strömbäck 2005; Wahl-Jorgensen 2001). Another fundamental value that hate speech undermines is equality, as it has the capacity to generate such a hostile environment that certain groups are marginalized from participating equally in the democratic aspects of society (George 2014).

By allowing such hate discourses to go unchallenged, journalism risks undermining its broader role as an authorized meaning-maker in public life (Johnson et al. 2021). This is also shown in the opinion of the professionals surveyed in Spain (Asociación de la Prensa de Madrid 2022), who believe that the reasons for the lack of public trust in journalism are sensationalism and the spectacle of the profession (57%), lack of rigor and information quality (45%), and lack of media independence and objectivity (37%).

The advent of the Internet and social media have prompted inquiries regarding the enduring relevance of the two-step flow of information hypothesis –by which information filters through journalists and the media before reaching the public–, but studies such as the one conducted by Alexandre et al. (2022) reaffirms that news organizations and journalists maintain their significance as communication agents.

Twitter is a social network that attracts users who seek social consumption of news (Tejedor et al. 2018), so its capacity to influence the media's agenda –with trending topics acting as a criterion for newsworthiness (López-Meri 2015)– is one of the main factors that explain why Twitter is a favored platform to stay informed for those under the age of 35 (Newman et al. 2023). Moreover, in contrast to other social media, Twitter presents a high concentration of political journalists (Lawrence et al. 2014; McGregor 2019), which may be due to its public-facing communication. Others, such as Facebook, tend to focus on more private communication (Tenenboim & Kligler-Vilenchik 2020). In addition, research such as Broersma & Graham (2016) show the impact of Twitter on journalists' work routines. Today, news professionals are looking for sources online, but their social networks are almost entirely composed of other journalists (Molyneux & Mourão 2017), making Twitter a

quasi-journalistic space.

For journalists, this platform has become, in addition, a professional tool that allows them to gain acceptance from the audience, not only for their work, but also for their personality (Molyneux & Holton 2014). They achieve this through a "discursive construction" of their identity (Olausson 2017), consisting of opinions, critique, ideological stances and emotional statements. They also share personal matters and even adopt a more informal and ironic tone to humanize themselves to their followers (López-Meri & Casero-Ripollés 2017; Cozma & Chen 2013; Lasorsa 2012; Sheffer & Schultz 2010; Vis 2013; Shifman 2007). Previous studies argue that voicing opinions and emotions, in contrast to traditional journalistic impartiality, brings the journalist closer to their audience (Cozma & Chen 2013; Lasorsa et al. 2012). Among journalists' self-promotion tactics, it has also been shown that tweeting during peak hours on current topics is useful (Brems et al. 2017).

Hedman and Djerf-Pierre (2013) categorize those whose professional attitudes are oriented toward adapting to audience demands and shaping their personal brand as "enthusiastic activists". This profile is often adopted by media directors, journalists who participate in television or radio discussions, columnists, and freelancers (López-Meri & Casero-Ripollés 2017). Among the reasons motivating these professionals to work on their personal image, Molyneux (2015, p. 931) identifies two motives: "a capitalistic endeavor whereby they become an influential voice regardless of their news organization", or "a narcissistic decision fueled by a simple human desire for attention".

Taking these demands of journalism into account, we ask these research question:

RQ3: Do positive or negative polarization prevail in the tweets of journalists published in a context of violence? What type of language do they use in their tweets?

RQ4: Amidst the hate campaign, what does the Twittersphere express about the role of journalists and media companies?

RQ5: Do users share journalistic content during these hate speech campaigns on Twitter? For what purpose do they share it?

Materials

Dataset acquisition and description

The dataset consists of 63,926 tweets ("Montero60K") containing several terms and hashtags related to "Irene Montero", collected between November 23 and November 27, 2022. Initially, we employed our tool Guaita (Pla et al. 2022), which leverages Twitter's API to collect real-time tweets on the subject. Subsequently, we conducted another download to ensure the data's accuracy by updating information on likes and retweets. This timeframe coincided with heightened political hostility towards the Spanish Minister, as elucidated in the introduction. Drawing from insights in academic literature (Congosto 2015; Suau-Gomila et al. 2020; Sánchez-Meza et al. 2023), we focused on the first 72 hours after a controversy, which is noted for the peak virality of social network content. The tweets were gathered following the verbal attack by far-right Vox deputy Carla Toscano on the Minister within the Spanish Parliament. This event was deemed significant for study purposes, considering the attack took place within one of the primary bodies of citizen representation, supposed to uphold a role model status. Moreover, Carla Toscano's intervention occurred during the debate on the 2023 General State Budget Bill on November 23rd, a crucial session in any legislative term.

A first subset of 2,239 tweets was selected based on the specific criteria that each message had garnered more than 50 likes or retweets during the specified period ("Montero2K" dataset). This subset collects the tweets with relative popularity attending the criterion that medium popularity of a tweet could be determined by one of these two ratios: resonance (number of likes generated ≥50) and/or virality (number of retweets generated ≥50) (Pancer & Poole 2016; Chen & Tan 2018; Oliveira et al. 2020). Within this group of relevance due to its diffusion, we generated a sub-sample with a special relationship with journalists. This sub-sample included tweets meeting one of three criteria: its author was an active journalist with more than 10,000 followers on Twitter (following the cut-off set by Alexandre et al. (2022)) and had to be employed by or collaborate regularly with a media outlet; its author was a reputable mass media, with more than 100,000 followers on the social network under study; its author was an individual who has embedded a hyperlink to a news article produced by a media outlet. As a result, we obtained a sub-sample of 413 tweets, which will serve as the test partition of our experiments.

Table 1 presents the statistics of the dataset.[1] Human experts manually labeled these tweets with "hate" and "improper language" categories. Furthermore, the 413 tweets in the test underwent extensive manual analysis by the human experts. Several examples of the analyzed tweets are provided in Table 2 (see examples in the first column). Figure 1 visually represents the word cloud generated from the Montero2K dataset text to illustrate word frequency, with the exclusion of search terms ("Irene" and "Montero"). Lastly, Figure 2 illustrates the presence of media accounts among the top 20 most-liked tweets.

Table 1. Statistics for the dataset and each partition for words. Number of tweets, number of words, average number of words per tweet (AWT), and vocabulary size (number of different words) in each of the three partitions of the selected Montero tweets (training, validation and test sets) composed of 2,239 tweets. The last row includes statistics for the entire acquired dataset.

Dataset	Partition	#Tweets	#Words	AWT	#Diff. words
Montero2K		2,239	83,739	37.40	10,332

	Training set	1,460	55,066	37.72	8,097
	Validation set	366	13,554	37.03	3,181
	Test set	413	15,119	36.61	2,961
Montero60K		63,926	2,805,321	43.88	22,794

Table 2. Examples of tweets from the test set. Tweets have been translated from Spanish and the original links are provided in the endnotes. Second and third columns show the given label by human experts.

Tweet	Hate	Improper
a) It is false that Montero is being attacked because she is a woman: because she is a woman she is trying to shield herself from criticism and avoid resignation after the yes is yes scandal. Three days she's been a victim on TV for an insult, just when rape victims were pointing the finger at her.[2]	Yes	No
b) Not everything is fair game @estherpalomera? You wrote 4 days ago in @eldiario.es that Irene Montero is bossed by "the alpha male" and you permanently whitewash Ana Rosa's media rubbish. Fascism advances also thanks to hypocrites like you.[3]	Yes	Yes
c) I find the words against Irene Montero unacceptable. No one should ever be attacked for being married to, for being the child of, for being the mother or father of. Never. It's the most despicable thing there is. And furthermore, in the case of "the wife of" it oozes machismo [4]	No	Yes
d) Serious question: do you think that Irene Montero, if she had not been with Pablo Iglesias, would also have become a minister? We read you! [5]	No	No
e) Well, of course it was not only yesterday in Congress, this bunch of shameless pigs and bitches also participated in the harassment of Irene Montero, feminism and decency. And there are a lot of names missing. And do not offend us with understanding. https://t.co/Wd5ACJv1Dj.[6]	Yes	Yes

Figure 1. Word cloud of the 100 most frequent words in Montero2K dataset.

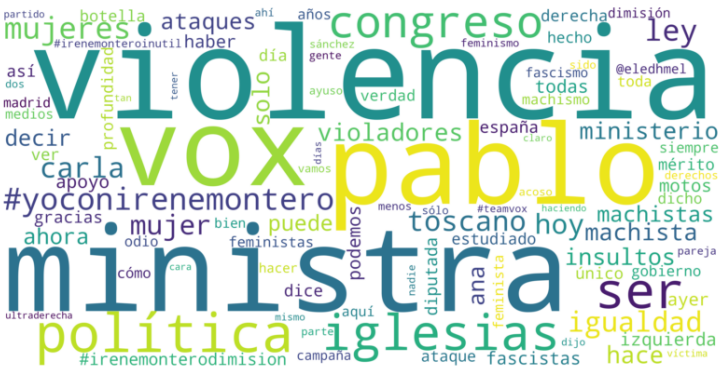
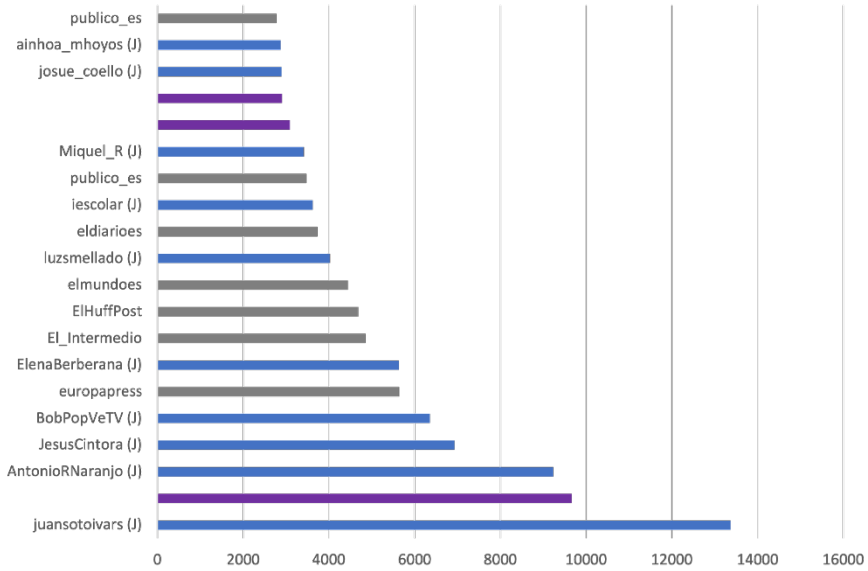


Figure 2. Presence of media and journalist accounts in the top 20 most-liked tweets from Montero2K dataset. Journalists are marked as "J". Untagged tweets belong to non-media accounts.



Human expert labeling process

Two expert journalists in the field labeled all entries (Montero2K dataset) within the selected dataset, for "hate" and "improper language" variables, indicating their presence or absence in the tweet. Moreover, a more detailed and qualitative content analysis was carried out in the sub-sample, individually assigning predefined categories to each entry. To do this, they configured a codebook made up of 5 main categorical variables and 6 optional variables, depending on authorship. The majority were inductively created by the authors after prior observation of the tweets; some deductively, derived from prior literature and existing codebooks.

1) The target of the attack:

- Irene Montero as she is the focus of our study
- The Ministry of Equality as a whole because there are people who consider that there is no need for it
- The Government, mostly because they "allow" Irene Montero to stay in it
- Podemos as her political party
- Feminism, with the same reason as the Ministry of Equality
- A specific journalist, either because it is supportive of the interlocutor's opinion, or because it is opposed to it
- Media, with the same reason as a specific journalist, but as a whole system
- Carla Toscano, the deputy who said the controversial statement
- Vox/Right as an extension to Carla Toscano
- Other
- Not applicable

2) The language used:

- Respectful and neutral: may express criticism, but in a constructive way
- Harsh: may use capital letters and is discrediting
- Harassing: includes insulting and denigrating expressions that enter the realm of siege (Tortajada et al. 2014)

3) The polarization towards Irene Montero. We previously defined polarization as the existence of ideological affinity between the political actor and the participants (Valera-Ordaz 2017):

- Favorable (positive) to Irene Montero or the target of the attack
- Contrary (negative) to Irene Montero or the target of the attack
- Neutral: there is polarization in favor and against and they somehow neutralize each other
- Null: does not apply

4) The authorship of the tweet:

- Media/programme/agency
- Journalist

- Media commentator, defined as an individual outside that particular media outlet, but who is invited to give his or her opinion on a subject
- Politician
- Citizen

5) Hate speech strategies (Martínez-Sanz et al. 2024; Noriega & Iribarren 2014; Tortajada et al. 2014)
- dehumanisation
- unverifiable assertion/misinformation
- divisive construction ("us-versus-them" dichotomy)
- incitement to violence
- insult or offensive term
- irony
- stereotype/prejudice promotion
- various strategies
- not applicable

In addition, depending on the authorship, we observed other variables. One of them was the type of media involved (this characteristic can influence the dissemination of the message, especially those that allow the circulation of photographs, illustrations and videos): print, radio television, podcast, news agency, institutional, fact-checker, cartoon, photograph or not applied.

In addition, it was measured whether a media outlet or journalist was mentioned in the tweet, as well as whether the tweet included a URL, image, video or quoted tweet. This feature might be used to spread a message and allow the user to comment on it.

At last, we included a variable on whether non-journalistic actors (politicians, citizens) used the media to reinforce their position, criticize its coverage, both, or neither (does not apply). This observation may be relevant because it provides a check on whether the media are being used to reinforce hate speech, even if they do not intend to do so, and what can be done about it. For this reason, we also measured whether the media outlet or journalist involved promoted hate speech. The categories are divided into yes, no, from another person or not applicable.

After this analysis, we have carried out an examination of the news pieces shared in the sample as they are the ultimate information in which hate speech is found or which may encourage it. Links to news items were found in 232 of the 413 tweets. With this number alone, it is easy to determine the importance they hold in the social network's messages, as more than half of the tweets include a news' link. The exploration of links to news items allowed us to detect the repetition of items. In other words, a particular news item had been viral enough to appear in several of the Twitter messages. At the end we compiled 138 unique articles which were subjected to the same rigorous analysis by two researchers as the other sample. We categorized them as the following:

- URL, to identify the news piece
- Date, to find out on which day there is a greater concentration of news
- Repeated, to get the most popular news in our samples
- Media, to identify the outlet
- Journalist, to check whether these news items are signed with name or only by the media outlet to which they belong
- Journalistic genre, we are interested in whether opinion articles or news items have been disseminated
- Polarization, as in the previous analysis, helps us to determine the position against, in favor, neutral or null regarding Irene Montero.
- Includes hate speech or not
- Where there is hate speech, by which we try to find out whether the hate speech is in quotation marks in a statement and the journalist only reports it or whether it is in the text written by the journalist him/herself
- Hate speech strategies, as in the previous analysis
- Target of attack, as in the previous analysis
- Location of hate speech, where can we find the hate speech: in the headline, in the pretitle or subtitle, in the text, in the audiovisual content, or in several of the above
- The specific topic of hate speech. The most common one was "Male chauvinist on Irene Montero"
- Hate speech in the image
- Types of images, differentiating whether it is an image of Irene Montero or of another politician, a screenshot from social media or another type of file
- Including video

Human expert labeling process for NLP

As previously stated, one of main focuses of this study was on the variables of "hate speech" and "improper language" (or "offensive terms"). We concentrated on these variables to perform automatic detection by using NLP techniques. The Montero2K dataset (2,239 tweets) was labeled with these categories (presence or absence of "hate speech" and "improper language"). The two expert human annotators labeled each tweet accordingly. The agreement between the annotators was

very high: in 99% of cases both annotators provided identical labels for the two categories, with a Kappa score of 0.98, indicating high agreement. In cases of disagreement, the two annotators engaged in discussions to reach a final consensus on the label. The annotation process resulted in a "ground truth" (also known as "gold standard") label for each category and tweet.

Table 2 shows several examples of tweets with their ground truth for the "hate" and the "improper" variables (second and third columns). In Table 3, the statistics for both classes in the whole dataset and the three partitions (training, validation, and test) are presented.

Table 3. Label distribution for "hate" and "improper" classes per partition of Montero2K dataset.

	Total	Hate	No hate	Improper	No improper
Training set	1,460	832 (56.99%)	628 (43.01%)	849 (58.15%)	611 (41.85%)
Validation set	366	208 (56.83%)	158 (43.17%)	213 (58.20%)	153 (41.80%)
Test set	413	241 (58.35%)	172 (41.65%)	243 (58.84%)	170 (41.16%)
Montero2K	2,239	1,281 (57.21%)	958 (42.79%)	1,305 (58.29%)	934 (41.71%)

Results from Natural Language Processing

Pretrained and fine-tuning NLP models

Pretrained Large Language Models (LLM) are neural language models that have been trained on large datasets before being fine-tuned or used for specific NLP tasks. These models learn contextualized representations of language during the initial pretraining phase, and they can then be fine-tuned for specific downstream tasks.

Currently, several pretrained LLMs are available for Spanish. Among these, MarIA (Gutiérrez-Fandiño et al. 2022) consists of ROBERTA and GPT-2 models trained on a corpus of 570GB of deduplicated texts with 135 billion words, sourced by the National Library of Spain between 2009 and 2019. BETO (Cañete et al. 2020), based on the BERT architecture with dynamic masking, was pretrained using datasets that include the Spanish Wikipedia, the Spanish portion of EUBookshop, and TED Talks. Both MarIA and BETO have demonstrated robust performance when fine-tuned for Spanish NLP tasks.

However, the specific characteristics of our task necessitated the consideration of alternative models. Our dataset comprises tweets, which are typically informal and often include Twitter-specific jargon. Additionally, the maximum input sequence length of MarIA and BETO is 512 tokens, which is significantly larger than the average tweet length (fewer than 50 tokens as shown in Table 4).

Table 4. Statistics for the dataset and each partition for tokens. Number of tweets, number of tokens, average number of tokens per tweet (ATT), and number of different tokens in each of the three partitions of the selected Montero tweets (training, validation and test sets) composed of 2,239 tweets.

Dataset	Partition	#Tweets	#Tokens	ATT	#Diff. tokens
Montero2K		2,239	108,697	48.55	10,373
	Training set	1,460	72,181	49.44	8,771
	Validation set	366	17,878	48.85	4,059
	Test set	413	18,638	45.13	3,815

To address these needs, we considered models specifically trained for the Twitter domain in Spanish. Two prominent models are TwilBERT and RoBERTuito. TwilBERT (González et al. 2021) was the first model trained using tweets in Spanish. It was developed using a dataset of 94 million tweet pairs in Spanish, with an adaptation of the BERT Next Sentence Prediction signal to the Twitter domain, termed Reply Order Prediction.

RoBERTuito (Pérez et al. 2022) is another LLM trained on a dataset of over 500 million tweets in Spanish, along with a small number of tweets in other languages such as Portuguese and English. RoBERTuito is available in three versions based on the preprocessing of the text: RoBERTuitoc (*case*d keeping the case found in the original tweets), RoBERTuitou (*uncased* version), and RoBERTuitod (*deaccente*d version which lower-cases and removes accents on tweets). Benchmark results published by the authors show that RoBERTuito's domain-specific models outperform general-purpose language models in various Spanish classification tasks (Domenech et al. 2023; Pérez et al. 2022). Moreover, RoBERTuito's maximum input sequence length of 128 tokens aligns well with the typical length of tweets, making it an ideal choice for our task.

Given these considerations, we opted to fine-tune the RoBERTuito models for our task. We trained the models using our dataset of 1,460 tweets and utilized a validation set to select the best models in terms of the evaluation metrics. We specifically employed Precision, Recall, and F1 Score as metrics for binary classification (Hate vs. No hate; and Improper

vs. No improper), along with Macro-Averaging (an average of Precision, Recall, and F1 Score across both classes), and Accuracy to measure overall prediction correctness (ratio of correctly classified instances to the total number of instances).

Table 5 presents the classification performance metrics for the hate and improper categories, indicating the performance of the best model on the validation set.

Table 5. Classification performance metrics for Hate and Improper categories for the best model on the validation set. Precision, Recall, and F1 score for both the No hate/Hate and No improper/Improper categories are given, as well as the Macro-Averaging of these measures for both classes; and, finally, the Accuracy of the model.

	Validation set					Validation set			
	Precision	Recall	F1 Score	Acc.		Precision	Recall	F1 Score	Acc.
No hate	0.786	0.740	0.762		No improper	0.856	0.864	0.860	
Hate	0.682	0.734	0.707		Improper	0.808	0.797	0.803	
Macro-Averaging	0.734	0.737	0.735		Macro-averaging	0.832	0.831	0.831	
Model				73.80%	Model				83.60%

NLP results for the test set

Figure 3 offers a comprehensive overview of the classification performance metrics and confusion matrices for the hate speech and improper language detection tasks on the test set.

The results reveal nuanced outcomes in the detection of hate speech and improper language. For the hate speech category, the model shows solid Precision (0.760) and Recall (0.867) for tweets labeled as "No hate", yielding an overall F1 Score of 0.810. However, the model faces challenges in identifying tweets labeled as "Hate", with adequate Precision (0.768) but lower Recall (0.616), leading to an F1 Score of 0.684. The confusion matrix also reveals that the model performs well in identifying tweets labeled as "No hate", achieving an accuracy of 87%. However, it faces challenges in correctly classifying tweets labeled as "Hate", with a lower accuracy of 62%. This suggests that the model struggles more when it comes to accurately identifying instances of hate speech.

In the case of improper language, the model excels in identifying tweets without improper language, as indicated by high Precision (0.731) and Recall (0.938), contributing to a nice F1 Score of 0.822. However, recognizing tweets containing improper language proves more demanding, with high Precision (0.851) but poor Recall (0.506), resulting in an F1 Score of 0.635. The confusion matrix highlights strong performance in correctly identifying tweets without improper language, with an accuracy of 94%. However, the model faces challenges in accurately classifying tweets containing improper language, with a poor accuracy of 51%. This suggests limitations in the model's ability to effectively detect instances of improper language in tweets.

Macro-Averaging provides a balanced assessment, yielding an average F1 Score of 0.747 for hate speech and 0.728 for improper language. The overall accuracy for both tasks stands at 76.27% and 76.03%, respectively.

Conclusively, the fine-tuned RoBERTuito models demonstrate promising capabilities in identifying hate speech and improper language in Spanish tweets, with notable strengths in certain categories and areas for improvement in others. The nuanced performance metrics provide insights into the models' effectiveness, and highlight avenues for future refinement, emphasizing the importance of continual evaluation and adaptation in the dynamic landscape of online content moderation.

To provide insights into the errors made by our model, the model misclassified tweets a) and e) from Table 2 within the Hate category and Improper language, respectively. In contrast, the model correctly categorized the rest of examples within both categories.

	Test set					Test set			
	Precision	Recall	F1 Score	Acc.		Precision	Recall	F1 Score	Acc.
No hate	0.760	0.867	0.810		No improper	0.731	0.938	0.822	
Hate	0.768	0.616	0.684		Improper	0.851	0.506	0.635	
Macro-Averaging	0.764	0.742	0.747		Macro-averaging	0.791	0.722	0.728	
Model				76.27%	Model				76.03%

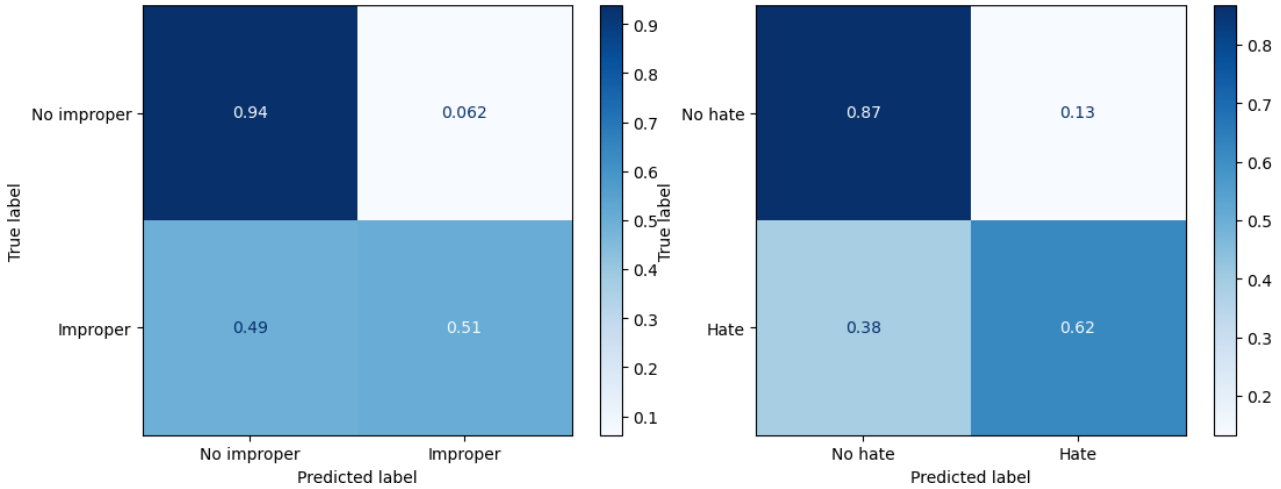


Figure 3. Classification performance metrics for the hate speech and improper language detection tasks on the test set. Precision, Recall, and F1 score for both the "No hate/Hate" and "No improper/Improper" categories are given, as well as the Macro-Averaging of these measures for both classes; along with the Accuracy of the model. Additionally, confusion matrices are shown for each category.

NLP analysis for Montero60K

Finally, while a quantifiable evaluation of the obtained models for the whole acquired dataset Montero60K (comprising 63,926 tweets) is unfeasible due to the absence of a ground truth for those tweets, a qualitative analysis can be conducted by specialized journalists, assuming that the model's output carries the errors identified in the previous results. The detected categories are shown in Table 6 for the 63,926 tweets.

Given the results on the Montero2K test and considering the Recall of the "No hate" class (0.867), the model demonstrates strong accuracy in correctly identifying tweets labeled as "No hate". However, tweets containing hate speech pose more challenges, with only 62% accurately identified (see Figure 3). These results suggest a high confidence in labeling tweets without hate by our model, indicating that the actual number of tweets with hate might be much higher than those detected. Considering this, the detection serves as a "lower bound", implying that at least 50.36% of the tweets may contain hate speech.

A similar scenario applies to the "Improper" class, where the majority of tweets labeled as "No improper" are accurately identified by the model (a Recall of 0.938 for the "No improper" class was achieved for the test set). This suggests that our model excels at detecting tweets without improper language, and considering this, at least 44.88% of the tweets may contain improper language, serving this figure as a "lower bound" in this context.

Table 6. Detected labels with NLP models on Montero60K (63,926 tweets).

	Automatic label detection	
	Total tweets detected	(%)
No hate	31,733	(49.64%)
Hate	32,193	(50.36%)
Total	63,926	(100%)
No improper	35,238	(55.12%)
Improper	28,688	(44.88%)
Total	63,926	(100%)

In Table 6 and answering **RQ1**, the results showcase a nearly balanced distribution between tweets categorized as No hate and Hate, with Hate slightly edging out. Similarly, there is a prevalence of tweets categorized as No improper compared to those categorized as Improper. The proportions indicate a relatively equal identification of both hate speech and improper language within the acquired dataset.

Conclusively, the qualitative assessment offers valuable insights into the model's performance across a vast dataset. The balanced distribution suggests a reasonable ability to identify hate speech and improper language. However, it is crucial to interpret these results with caution, recognizing potential limitations and the need for ongoing refinement.

Results from Qualitative Analysis

Journalists inside polarization

Answering to **RQ2**, in Montero2K, 150 tweets were published by 83 journalists with more than 10,000 followers.

Statistically the mode is 1 tweet during the period of 3 days. The most active was @yagoalons (11 tweets), a journalist of the conservative media VozPópuli. He is followed by @ldpsincomplejos (8 tweets), a senior position in conservative radio esRadio, and @rosamariaartal (7 tweets), a columnist of the left-wing digital native newspaper eldiario.es. In their profile, 76.19% identify the media they work for, almost half conservative and half progressive in ideology. Regarding their professional profile, 38.10% of these journalists are reporters, 23.81% are columnists, and 20.24% are media directors (see Table 7). It is important to differentiate that 8 are legacy and native media directors, as well as 9 are pseudo-media managers. Regarding gender, 56 journalists are men and 27 are women.

Table 7. Professional roles of journalists and their use of hate speech.

Professional role	Unique identities	Not hate speech	Hate speech	Hate speech from another person	Tweets number
Director of a media outlet	8	10	2	1	13
Director of a pseudo-media	9	10	7	2	19
Management position	8	11	7	3	21
Management position in a pseudo-media	3	3	2		5
Journalist	32	23	19	8	50
Columnist	20	20	14	2	36
Photojournalist	1	1			1
Cartoonist	2		4		4
Radio programme	1			1	1
Total general	84	71	55	17	150

When the journalists' attitude towards online misogyny on Twitter is examined, 37.58% (N=56) of their tweets incorporate forms of hate speech, 11.41% (17 tweets) include quotes with hate speech and 51.01% do not present any of the types of hate speech studied. The most frequent users of hate speech -used in favor and against online misogyny campaigns- are reporters (33.93%) and columnists (26.79%). It is noteworthy that the type of hate speech on journalists' accounts varies depending on their ideological stance regarding the hate campaign against Montero, which already advances a response to **RQ3**. When these professionals defend the Minister from the hate campaign she is facing, they resort to insults and offensive terms. For example, they label those who insult Montero as "fascists", "Nazi horde", "rabid dying Francoists", "unleashed coven" or "bunch of shameless pigs and bitches" (see example number 5 on Data and Methods or https://bit.ly/47zILvT). Their target of attack is the far-right party Vox and/or its deputy Carla Toscano.

In contrast, those who support the arguments of the campaign against the Minister often use divisive language and the "us-versus-them" dichotomy. On one hand, they shift the meaning of "violence" against Montero to what she has supposedly generated with the approval of the "Only Yes Means Yes" law (https://bit.ly/47WU01n) and the subsequent early release of some rapists. On the other hand, they justify the hate campaign by pointing out the alleged mistreatment that the Minister and her party have exerted against the opposing ideological side (https://bit.ly/47CSrpt). On a second level and following this line, Irene Montero is the target of the attack in 80% of the cases, against whom accusations are leveled based on metaphors and stereotypes around gender roles. Much less frequently, the offenses are directed against her political party (2.5%), the left governing coalition (2.5%) or the ideological left (2.5%). Irony is employed to emphasize the Minister's management mistakes (https://bit.ly/3N2tv2I).

When observing the 20 tweets with the greatest dissemination in the Montero2K sample (see Figure 2), half of them were published by journalists and 7 media outlets, which shows the communicative potential of content linked to the profession. Moreover, if we look at the circulation figures of tweets by journalists that contain hate speech or a quote that includes hate speech, 47 tweets reached at least 100 likes and 28 were retweeted at least 100 times. The two tweets with the highest number of likes (13,367 and 9,232, respectively) and retweets (4,544 and 3,252, respectively) take a stance against Irene Montero by using a divisive construction. In this case study, users tend to express their support primarily through likes rather than retweets.

In addition to the 150 tweets published by journalists, there were 168 tweets posted by 42 media accounts (see Table 8). Of these, 70.24% (118 tweets) do not display hate speech, while 27.38% (48 tweets) include quotations containing hate speech. Four of the tweets posted by media outlets (2.38%) do incorporate some form of hate speech in their text: three tweets are conservative digital media (El Confidencial, Libertad Digital and The Objective) and one is a left-wing podcast led by the ex politician and Podemos' founder Pablo Iglesias (La Base). Again, the communicative strategy of these varies based on ideological positioning. The only tweet expressing support for Irene Montero among these 4 refers to her aggressors using an offensive term ("los ultras") as seen in https://bit.ly/47tHmXT. On the other hand, those opposing the Minister (3 tweets) employ irony and foster the prejudice that she is a theatrical/martyr woman, using the fact that she cried after Carla Toscano's words to divert attention from the controversy surrounding the new law regarding sexual consent. They also use this argument to reinforce the stereotype that tears or showing emotions are feminine traits and signs of weakness associated with women politicians (Jones 2016). Except for the tweet in support of Montero (with 355 retweets and 220 likes), the engagement for the other three tweets was limited.

In the case we are analyzing, it is relevant to look at the 48 tweets that reproduce quotes including hate speech because they

may be reinforcing the ongoing hostility. These are citations in quotation marks or video fragments. 31.25% (15 tweets) include an insult or offensive term ("violence worshiper", "satrap", "gang of fascists"), and 29.17% (14 tweets) reproduce the stereotype of Montero as a woman who has benefited from a man. We would like to emphasize that more than one third of these tweets -35.42% (17 tweets)- include more than one typology of hate speech, with the most frequent being insults (14 tweets), stereotype promotion (8 tweets), divisive construction (7 tweets), unverifiable assertion/misinformation (3 tweets) and dehumanization (2 tweets). This leads us to think about the complexity and virulence of hate speech.

Table 8. Profile of tweet creators linked to journalistic information.

Profile	Number of tweets
Citizenship (mentioning media/journalist or sharing media URL)	66
Journalist	150
Media/News Agency	168
Media opinion leader (not journalist/politician)	11
Politician	18
Total	413

Social and ethical responsibility of journalists

On the prevalence of positive or negative polarization (**RQ3**), 88% of the journalists' tweets (150) take a stance either in favor or against Irene Montero; 10% adopt a critical position (expressing both support and opposition), and 3% remain neutral (Table 9). In other words, journalists are virtually divided between the two extremes of opinion.

Table 9. Journalists' polarization and hate.

Polarization	No hate	Hate	Hate speech from another person	Total
Opposite	17	35	4	56
Favorable	45	20	11	76
Neutral	3			3
Null	13		2	15
Total	78	55	17	150

When we examine the language they use, 52.67% (79 tweets) express themselves with respectful and neutral language; 34.67% (52 tweets) employ harsh language (https://bit.ly/3R5SK5o); 11.33% reproduce harsh language from someone else -encapsulated in quotation marks- and 1.33% express themselves in a extreme harassing manner (https://bit.ly/3R5SK5o).

It is striking that the social responsibility of the media in these hate campaigns also generates a reflection when we observe the sample of terms made up of the most widely circulated tweets (Figure 1). When looking at the 100 most used words by the Twittersphere (Figure 1), "media" appears in position 66 ($N = 56$) and the surname of a television journalist "Motos" is detected in position 34 ($N = 90$), in whose programme the Minister's policies were questioned. It should also be noted that among the most repeated terms are the two ideological poles of the debate: "left" ($N = 94$) and "right" ($N = 75$).

Without taking into consideration the 150 tweets published by journalists and the 168 tweets published by media companies, 10.33% of the 2,233 tweets published by other users focus on the professional practice of journalists (RQ4). These 198 tweets point to the extreme polarization of the media and journalists, and blame them for provoking these hate campaigns or silence attacks on political parties with which their users do not sympathize. Expressions such as "this miserable hatred is sown by the media", "the media are directly responsible for the political violence" and "this is the tip of the iceberg of a violence that extends to other places: social networks, the media and all the media loudspeakers" are evidence of the unease expressed by a part of the public opinion with the lack of ethics in Spanish journalism.

Media content as an argument for hate speech

To answer **RQ5**, we have proceeded to the analysis of media links shared by users in Twitter and present in 339 tweets. They have been shared by their own media (166 tweets), journalists (84 tweets), citizens (64 tweets), politicians (18 tweets) and columnists (7 tweets). In the case of citizens, 75% of the time they share a link to a piece of news to reinforce their position, and 12.5% to criticize that piece of news coverage. It is not possible to discern its use on 7 occasions, and on one occasion they both applaud and criticize a piece of news at the same time. When the user is a politician, 72.22% of the time the link to a media outlet reinforces their argument.

In the entire sample analyzed, 138 unique pieces of journalism have been shared, the majority signed by the progressive digital native eldiario.es (20 news items). It was followed by another native left-wing newspaper, Público, with 10; the conservative native Libertad Digital with 9; the private news agency Europa Press with 8; and neoliberal digital native newspaper Okdiario with 7. The day on which the most news about the event was published was 24 November, not the 23

itself; so we can consider that day as the peak of the controversy.

As for the authors, we can state that in 30 news items the author cannot be identified (21.74%). Others do not even come close, such as Europa Press (8) or eldiario.es (4), which are also the signature of the company and not of a specific journalist. This leads us to think that these pieces of news are not backed up by the work of a journalist, but are news items heavily based on agency teletypes.

In the analysis of the journalistic genre, the majority of the pieces are news, with 91 articles (65.94%). Others of lesser relevance are opinion articles (22), chronicles (12) or interviews (2). It would be expected then that, being informative pieces, they would not be positioned on any side of the political spectrum. We can confidently claim that this has not been the case. Although there are 65 pieces (47.10%) that do not contain any polarization, the sample is divided: 32 of the pieces are positioned against Irene Montero, 38 in favor and 3 are neutralized (they contain arguments for and against).

On the contrary, we find a very clear answer to whether these news items include hate speech: 76.81% do. From these, we can extract that in 50% of them (69 pieces) this hate speech is found in statements in quotation marks, while in 23.19% this speech is written by the journalist in their own text. As for the type of hate speech used, it can be seen that in 73 of the journalistic pieces (52.90%) there are several types of hate speech in the same tweet, which, as said, implies more virulence in the attack. Thus, the most repeated types of hate speech are the insult or offensive term in 89 of the news pieces (64.49%), the message spreads or promotes a stereotype in 57 (41.30%) and encourages a divisive construction in 18 (13.04%). We did not find as much dehumanization as in the tweets. In this analysis, they only constitute 2.90% of the sample.

When analyzing the location of hate speech, we can affirm that 50.72% of it is found in the text of the news item to a greater extent, and only in 29 articles (21.01%) we see a preponderance of hate speech in the headline, contrary to the findings of Palau-Sampio and Carratalá (2022) in far-right wing Spanish pseudo-media, where clickbait headlines were predominant. Finally, we wanted to check whether there was hate speech in the images or videos included in the news. The use of videos was low, as only 21 news items included them (15.21%); and although all the news items had an image, mostly of Irene Montero, we only detected hate speech in 13 of them (9.42%). These images portrayed the Minister as either an aggressive woman trying to dominate the political scene by crying for pity, or as a frivolous person who spends taxpayers' money on nonsense.

Discussion and conclusions

Several academic studies have examined the presence of journalists on this microblogging network characterized by the spread of breaking news (Kovach & Rosenstiel 2010), but also controversial due to the dissemination of hate speech (Llorca-Asensi et al. 2021; Acosta-Quiroz & Iglesias-Osores 2020). On Twitter, media professionals have assumed a key role in shaping public opinion as a guide or mentor (Garcia & Marta-Lazo 2017). They have gained influence not only through their professional practice but also through their discursive personality (Molyneux & Holton 2014). Perceived as a trustworthy authority —most indicate their profession and the journalistic organizations they work for in their biographies— the novelty of this research lies in focusing attention on the weight of their emotions and opinions surpassing facts within the campaigns of polarization, harassment and, above all, hate speech generated on this network against women politicians (Carr & McCracken 2018). To this end, we first ensure that this case study takes place inside a gender hate polemic, as confirmed by both Natural Language Processing experts and coders specialized in this type of discourse.

This research contributes to the relationship between journalistic ethics and the use of information technologies on several aspects. So far, the research community has studied the attacks against journalists on this platform, especially women (Bhat 2024; Sampaio-Dias et al. 2024). But, on this occasion, the analysis puts the spotlight on the activity of the journalists themselves. It reaffirms their influential capacity, as tweets posted by journalists rank among the most widely disseminated in these campaigns. Moreover, the results are particularly concerning given the striking presence of various forms of hate speech in the tweets published by journalists as they exhibit the use of offensive terms, divisive constructions, irony and the promotion of stereotypes. Supported by a lax interpretation of the right to freedom of expression on this platform (Johnson et al. 2021), in half of the cases in which they intervene they limit their reporting to the repetition of hate speeches verbalized by others or even by themselves.

The study shows an extreme polarization among "influencer" journalists, unable to analyze the situation from the perspective of deliberative democracy. This work questions the journalists' social responsibility as critical observers of reality by revealing an ideological stance towards one of the two sides of the conflict -for and against the Minister's actions-. As Nikolaev et al. (2023) state, it is precisely their conviction of moral superiority that leads them to express negative emotions against their adversaries and justify their hate speech.

This professional attitude, based on freedom of speech, neglects the moral obligation to nurture democratic health by demanding ethical standards in conversation, minimizing harm and educating the public on the complexity and consequences of hate speech. For these reasons, self-regulation is necessary. The most paradoxical aspect is that, as we have seen, some journalists are combating hate speech with hate speech.

A similar situation arises with the media. Although they may not explicitly employ the discursive strategies of hate speech, they do function as echo chambers for such narratives (Blanco-Alfonso et al. 2022; Lacalle et al. 2023; Alexandre et al. 2022), relying on the quoting of a hate speech statement, usually placed in headlines and subheadings. At the same time, we have found that citizens and political leaders cited more frequently pieces that promote hate speech than those that do not,

which highlights the importance of the media in the defense of political positions.

This evidence calls for two types of ethical standards within the profession. On one hand, it is urgent to establish guidelines to regulate hate speech expressed by journalists on Twitter, although this measure can only be applied when their accounts are linked to a media outlet in their biography (most often). On the other hand, standards are needed to prevent the excessive dissemination of hate speech in journalistic pieces, both those verbalized by others and by journalists themselves. These requirements are considered particularly necessary in a context of professional discredit such as the one we are living in worldwide. In fact, this loss of credibility can even be observed in the conversation analyzed, where one out of ten tweets single out media and journalists, especially those related to television, for preparing the breeding ground for these hate groups.

The results of this study indicate promising outcomes, demonstrating the NLP models' competitive performance in hate and improper language detection, underscoring their potential for effective applications in content moderation and enhancing online safety. In this regard, future research should expand the range of sample collection, as well as include other women politicians at national, regional and local levels susceptible to hate speech for promoting gender politics, for whose analysis this article sets the foundation. As a limitation to the research, it should be noted that in a context of controversy like the one presented, tension is likely to be exacerbated and the community participating in the social conversation interacts conditioned by their displeasure or support for the words of the congresswoman.

Acknowledgements

This research is part of the projects BEWORD-UPV: PID2021-126061OB-C41 and MCIN/AEI: PID2020-113574RB-I00 (Disflows), as well as to the call for predoctoral contracts 2022 of the Universidad de Valladolid, co-funded by Banco Santander.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by two Spanish Ministry of Science and Innovation research projects (PID2020-113574RB-I00 and PID2021-126061OB-C41).

Notes

1. Dataset available at https://huggingface.co/datasets/ELiRF/montero2k.
2. https://twitter.com/juansotoivars/status/1596421946899824642.
3. https://twitter.com/PabloIglesias/status/1595691120813838336.
4. https://twitter.com/BeatrizTalegon/status/1595730563926822912.
5. https://twitter.com/MediterraneoDGT/status/1595825838716772352.
6. https://twitter.com/rosamariaartal/status/1595769542508412930.

References

Abuín-Vences N, Cuesta-Cambra U, Niño-González JI & Bengochea-González C (2022) Hate speech analysis as a function of ideology: Emotional and cognitive effects. *Comunicar* 30(71): 37–48. DOI:10.3916/C71-2022-03.

Acosta-Quiroz J & Iglesias-Osores S (2020) Covid-19: desinformación en redes sociales. *Revista cuerpo médico HNAA* 13(2): 217–218. DOI:10.35434/rcmhnaaa.2020.132.678.

Akbar M & Safdar A (2023) Politics of Hate and Social Media: Thematic Analysis of Political Hate Discourses on Facebook. *Global Social Sciences Review* 8(2): 364–375. DOI:10.31703/ gssr.2023(VIII-II).33.

Alexandre I, Jai-sung Yoo J & Murthy D (2022) Make Tweets Great Again: Who Are Opinion Leaders, and What Did They Tweet About Donald Trump? *Social Science Computer Review* 40(6): 1456–1477. DOI:10.1177/08944393211008859.

Alonso-González M (2019) Fake news: disinformation in the information society. *Ámbitos. Revista internacional de comunicación* 45: 29–52. DOI:10.12795/Ambitos.2019.i45.03.

Arencón Beltrán S, Morales S & Hernández Conde M (2023) Fascismo digital para bloquear la participación y la deliberación feminista. T*eknokultura. Revista de Cultura Digital y Movimientos Sociales* 20(1): 25–35. DOI:10.5209/TKN.81002.

Asociación de la Prensa de Madrid (2024) FAPE, APM y APP confirman que no han recibido ninguna solicitud de amparo por parte del PSOE. URL https://www.apmadrid.es/comunicado/fape-apm-y-app-confirman-que-no-han-recibido-ninguna-solicitud-de-amparo-por-parte-del-psoe/

Asociación de la Prensa de Madrid (2022) Informe de la Profesión Periodística 2021. Madrid: Asociación de la Prensa de Madrid.

Barberá P (2014) Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data. *Political Analysis* 23(1): 76–91. DOI:10.1093/pan/mpu011.

Bhat P (2024) Coping with Hate: Exploring Indian Journalists' Responses to Online Harassment. Journalism Practice 18(2): 337–355. DOI:10.1080/17512786.2023.2250761.

Blanco-Alfonso I, Rodríguez-Fernández L & Arce-García S (2022) Polarización y discurso de odio con sesgo de género asociado a la política: Análisis de las interacciones en Twitter. *Revista de Comunicación* 21(2): 33–50. DOI:10.26441/rc21. 2-2022-a2.

Brems C, Temmerman M, Graham T & Broersma M (2017) Personal Branding on Twitter: How employed and freelance journalists stage

themselves on social media. *Digital Journalism* 5(4): 443–459. DOI:10.1080/21670811.2016.1176534.

Broersma M & Graham T. (2016). Tipping the balance of power social media and the transfor-mation of political journalism. In: Bruns A, Enli G, Skogerbo E, et al. (eds) *The Routledge Companion to Social Media and Politics.* London: Routledge, pp. 89–103.

Bruns A (2021) Echo chambers? Filter bubbles? The misleading metaphors that obscure the real problem. In Marta Pérez-Escolar, José Manuel Noguera-Vivo (eds.) Hate Speech and Polarization in Participatory Society. Routledge. DOI:10.4324/9781003109891

Cabezas Fernández M, Pichel-Vázquez A & Enguix Grau B (2023) El marco "antigénero" y la (ultra)derecha española. Grupos de discusión con votantes de Vox y del Partido Popular. *Revista de Estudios Sociales* 85: Article 85. DOI:10.7440/ res85.2023.06.

Callado R (2021) "Why Can't We?" Disinformation and Right to Self-Determination. The Catalan Conflict on Twitter. Social Sciences 10(10): 1–23. DOI:10.3390/socsci10100383

Cañete J, Chaperon G, Fuentes R, Ho JH, Kang H and Pérez J (2020) Spanish pre-trained BERT model and evaluation data. In: *PML4DC at ICLR* 2020. DOI: 10.48550/arXiv.2308.02976

Carr A & McCracken H (2018) 'Did we create this monster?' How Twitter turned toxic. *Fast Company* URL https://www.fastcompany.com/40547818/did-we-create-this-monster-how-twitter-turned-toxic.

Chen F & Tan WH (2018) Marked self-exciting point process modelling of information diffusion on Twitter. The Annals of Applied Statistics, 12(4): 2175–2196. DOI:10.1214/18-AOAS1148

Cheng Z, Marcos-Marne H & Gil de Zúñiga H (2023) Birds of a Feather Get Angrier Together: Social Media News Use and the Social Media Political Homophily as Antecedents of Political Anger. *Political Behavior* DOI:10.1007/s11109-023-09864-z.

Cinelli M, Morales G, Galeazzi A, Quattrociocchi W & Starnini M (2021) The echo chamber effect on social media. *Proceedings of the National Academy of Sciences of the United States of America* 118(9): e2023301118. DOI:10.1073/pnas. 2023301118.

Cohen-Almagor R (2013) Freedom of Expression v. Social Responsibility: Holocaust Denial in Canada. *Journal of Mass Media Ethics* 28(1): 42–56. DOI:10.1080/08900523.2012.746119.

Colleoni E, Rozza A & Arvidsson A (2014) Echo chamber or public sphere? Predicting political orientation and measuring political homophily in Twitter using big data. *Journal of Communication* 64(2): 317–332. DOI:10.1111/jcom.12084.

Congosto M (2015) Elecciones Europeas 2014: Viralidad de los mensajes en Twitter. *Redes. Revista hispana para el análisis de redes sociales* 26(1): 23–52. DOI:10.5565/rev/redes.529.

Corbal M. (2022, November 26th) Carla Toscano, la diputada de Vox que ofendió a Montero: rock, Marvel, García Lorca, hijos y su gato Don Pelayo. El Mundo. URL https://www.elmundo.es/loc/famosos/2022/11/26/6380a495fdddffbe708b459d.html

Cozma R & Chen K (2013) What's in a tweet? Foreign correspondents' use of social media. *Journalism Practice* 7(1): 33–46. DOI:10.1080/17512786.2012.683340.

Del-Fresno-Garcí-a M (2019). Desórdenes informativos: sobreexpuestos e infrainformados en la era de la posverdad. *Profesional De La información* 28(3). DOI:10.3145/epi.2019.may.02

Domenech L, Pérez JM, Rosati G & Kozlowski D (2023) Gender biases and hate speech: promoters and targets in the Argentinean political context. SocArXiv DOI:10.31235/osf.io/6cts8

Durántez-Stolle P, Martínez-Sanz R, Piñeiro Otero T & Gómez-García S (2023) Feminism as a Polarizing Axis of the Political Conversation on Twitter: The Case of IreneMonteroDimision. *El Profesional de la Información* (e320607). DOI:10.3145/epi.2023.nov.07.

Figenschou T U & Ihlebaek K A (2019) Challenging journalistic authority: Media criticism in far-right alterna-tive media. *Journalism studies* 20(9), 1221–1237 DOI:10.1080/1461670X.2018.1500868

Flamino J, Galeazzi A, Feldman S, Macy MW, Cross B, Zhou Z, Serafino M, Bovet A & Makse HA (2023). Political polarization of news media and influencers on Twitter in the 2016 and 2020 US presidential elections. *Nat Hum Behav* 7, 904–916 DOI:10.1038/s41562-023-01550-8

Fletcher R, Cornia A & Nielsen RK (2020) How Polarized Are Online and Offline News Audiences? A Comparative Analysis of Twelve Countries Research Article. *The International Journal of Press/Politics* 25(2): 169–195. DOI:10.1177/1940161219892768.

Frenkel S & Conger K (December 2 2022) Hate Speech's Rise on Twitter Is Unprecedented, Researchers Find. The New York Times. URL https://www.nytimes.com/2022/12/02/technology/twitter-hate-speech.html

García M & Marta-Lazo C (2017) Análisis de Twitter como fuente, recurso de interacción y medio de difusión para los periodistas vascos. ZER. *Revista de Estudios de Comunicación* 22(42): 73–95. DOI:10.1387/zer.17833.

George C (2014) Journalism and the Politics of Hate: Charting Ethical Responses to Religious Intolerance. Journal of Mass Media Ethics 29(2): 74–90. DOI:10.1080/08900523.2014.893771.

Ging D & Siapera E (2019) *Gender Hate Online: Understanding the New Anti-Feminism*. Springer International Publishing.

González J, Hurtado LF & Pla F (2021) Twilbert: Pre-trained deep bidirectional transformers for Spanish Twitter. Neurocomputing 426: 58–69. DOI:10.1016/j.neucom.2020.09.078.

Gutiérrez-Fandiño A, Armengol-Estapé J, Pàmies M, Llop-Palao J, Silveira-Ocampo J, Carrino CP, Armentano-Oller C, Rodriguez-Penagos C, Gonzalez-Agirre A & Villegas M (2022) MarIA: Spanish Language Models. *Procesamiento del Lenguaje Natural* 68: 39–60. DOI:10.26342/2022-68-3.

Hawdon J, Costello M, Bernatzky C & Restifo SJ (2022) The Enthymemes of Supporting President Trump: Explaining the Association Between Structural Location, Supporting the President, and Agreeing With Online Extremism. *Social Science Computer Review* 40(1): 24–41. DOI:10.1177/0894439320905767.

Hawdon J, Reichelmann A, Costello M, Llorent VJ, Räsänen P, Oksanen A & Blaya C (2023) Measuring Hate: Does a Definition Affect Self-Reported Levels of Perpetration and Exposure to Online Hate in Surveys? *Social Science Computer Review* 0(0). DOI:10.1177/08944393231211270.

Hedman U & Djerf-Pierre M (2013) The Social Journalist: Embracing the social media life or creating a new digital divide? Digital

Journalism 1(3): 368–385. DOI:10.1080/21670811.2013.776804.
Hermida X (2022) Escándalo en el Congreso tras espetar Vox a Irene Montero: "Su único mérito es haber estudiado en profundidad a Pablo Iglesias". El País. URL https://elpais.com/espana/2022-11-23/vox-insulta-a-irene-montero-en-el-congreso-llamandola-libertadora-de-violadores.html
Hernández-Santaolalla, V. & Sola-Morales, S. (2019). Postverdad y discurso intimidatorio en Twitter durante el referéndum catalán del 1-O. Observatorio (OBS*), 13(1), 102–121. DOI:10.15847/obsOBS13120191356
Iranzo-Cabrera M & Casero-Ripollés A (2023) Political entrepreneurs in social media: Self-monitoring, authenticity and connective democracy. The case of Íñigo Errejón. *Heliyon* 9(2). DOI: 10.1016/j.heliyon.2023.e13262.
Johnson BG, Thomas RJ & Kelling K (2021) Boundaries of Hate: Ethical Implications of the Discursive Construction of Hate Speech in U.S. Opinion. *Journal of Media Ethics* 36(1): 20–35. DOI:10.1080/23736992.2020.1841643.
Jones JJ (2016) Talk "Like a Man": The Linguistic Styles of Hillary Clinton, 1992–2013. *Perspectives on Politics* 14(3): 625–642. DOI:10.1017/S1537592716001092.
Jones S (2022) Spanish right launch sexist attacks on equality minister over consent law. The Guardian. URL https://www.theguardian.com/world/2022/nov/24/irene-montero-spanish-right-sexist-attacks-consent-law
Jurkowitz M & Gottfried J (2022) Twitter is the go-to social media site for U.S. journalists, but not for the public. pew Research Center. URL https://www.pewresearch.org/short-reads/2022/06/27/twitter-is-the-go-to-social-media-site-for-u-s-journalists-but-not-for-the-public/
Kim JN & Gil d Zúñiga H (2021) Pseudo-Information, Media, Publics, and the Failing Marketplace of Ideas: Theory. American Behavioral ScientistVolume 65(2), 163–179. DOI:10.1177/0002764220950606
Kimmel M (2019) *Angry White Men: American Masculinity at the End of an Era*. New York: Nation Books.
Konikoff D (2021) Gatekeepers of toxicity: Reconceptualizing Twitter's abuse and hate speech policies. *Policy & Internet* DOI:10.1002/poi3.265.
Kovach B & Rosenstiel T (2007) *The Elements of Journalism: What Newspeople Should Know and the Public Should Expect.* New York: Random House.
Kovach B & Rosenstiel T (2010) *Blur: How to Know What's True in the Age of Information Overload.* New York: Bloomsbury Publishing.
Kreiss D & McGregor SC (2024) A review and provocation: On polarization and platforms. *New media & society* 26(1): 556–579. DOI:10.1177/14614448231161880
Lacalle C, Martín Jiménez V & Etura Hernández D (2023) El antifeminismo de la ultraderecha española en Twitter en torno al 8M. *Revista Prisma Social* 40: 358–376. URL https:// revistaprismasocial.es/article/view/4837.
Lasorsa DL (2012) Transparency and Other Journalistic Norms on Twitter. *Journalism Studies* 13(3): 402–417. DOI:10.1080/ 1461670X.2012.657909.
Lasorsa DL, Lewis SC & Holton AE (2012) Normalizing Twitter. Journalism Practice in an Emerging Communication Space. *Journalism Studies* 13(1): 19–36. DOI:10.1080/1461670X.2011.571825.
Lawrence RG, Molyneux L, Coddington M & Holton A. (2014). Tweeting Conventions: Political journalists' use of Twitter to cover the 2012 presidential campaign. Journalism Studies, 15(6): 789–806. DOI:10.1080/1461670X.2013.836378.
Leader-Maynard J & Benesch S (2016) Dangerous speech and dangerous ideology: An integrated model for monitoring and prevention. *Genocide Studies and Prevention: An International Journal* 9(3): 70–95. DOI:10.5038/1911-9933.9.3.1317.
Lingiardi V, Carone N, Semeraro G, Musto C, D'Amico M & Brena S (2020) Mapping Twitter hate speech towards social and sexual minorities: a lexicon based approach to semantic content analysis. *Behaviour & Information Technology* 39(7): 711–721. DOI:10.1080/0144929X.2019.1607903.
Llorca-Asensi E, Sánchez Díaz A, Fabregat-Cabrera ME & Ruiz-Callado R (2021) "Why Can't We?" Disinformation and Right to Self-Determination. The Catalan Conflict on Twitter. *Social Sciences* 10(10): 1–23. URL https://www.mdpi.com/2076-0760/10/10/383/pdf.
López-Meri A (2015) El impacto de Twitter en el periodismo: un estado de la cuestión. *Revista de la Asociación Española de Investigación de la Comunicación* 2(4): 34–41. URL http://www.revistaeic.eu/index.php/ raeic/article/view/55.
López-Meri A & Casero-Ripollés A (2017) Las estrategias de los periodistas para la construcción de marca personal en Twitter: posicionamiento, curación de contenidos, personalización y especialización. *Revista Mediterránea de Comunicación* 8(1): 59–73. DOI:10.14198/MEDCOM2017.8.1.5.
Martínez-Sanz R, Durántez-Stolle P, Simón-Astudillo I (2024). Memes as hate speech: Violence, humour and criticism around the image of Irene Montero. *VISUAL REVIEW. International Visual Culture Review Revista Internacional De Cultura Visual* 16(2), 1–16. DOI:10.62161/revvisual.v16.5193
Massanari A (2017) #Gamergate and The Fappening: How Reddit's Algorithm, Governance, and Culture Support Toxic Technocultures. *New Media & Society* 19(3): 329–346. DOI: 10.1177/1461444815608807.
Matuszewski P & Szabó G (2019) Are Echo Chambers Based on Partisanship? Twitter and Political Polarity in Poland and Hungary. Social Media + Society 5(2). DOI:https://doi.org/10.1177/2056305119837671
McGregor S (2019). Social media as public opinion: How journalists use social media to represent public opinion. *Journalism* 20(8), 1070-1086. DOI:10.1177/1464884919845458.
Michailidou A & Trenz HJ (2021) Rethinking Journalism Standards in the era of Post-Truth Politics: From Truth Keepers to Truth Mediators. *Media, Culture Society* 43(7): 1340–1349. DOI:10.1177/01634437211040669.
Milli S, Carroll M, Wang Y, Pandey S, Zhao S & Dragan A (Jan. 3, 2024) Engagement, User Satisfaction, and the Amplification of Divisive Content on Social Media. *Knight First Amend. Inst.* URL https://knightcolumbia.org/content/engagement-user-satisfaction-and-the-

amplification-of-divisive-content-on-social-media

Mills T, Mullan K & Fooks G (2021) Impartiality on Platforms: The Politics of BBC Journalists' Twitter Networks. *Journalism Studies* 22(1): 22–41. DOI:10.1080/1461670X.2020.1852099.

Miškolci J, Kováčová L & Rigová E (2020) Countering Hate Speech on Facebook: The Case of the Roma Minority in Slovakia. *Social Science Computer Review* 38(2): 128–146. DOI:10.1177/0894439318791786.

Molyneux L & Mourão RR (2019) Political Journalists' Normalization of Twitter: Interaction and New Affordances. *Journalism Studies* 20(2): 248–266. DOI:10.1080/1461670X.2017.1370978.

Molyneux L (2015) What Journalists Retweet: Opinion, Humor, and Brand Development on Twitter. *Journalism* 16(7): 920– 935. DOI:10.1177/1464884914550135.

Molyneux L & Holton A (2014) Branding (health) journalism: Perceptions, practices, and emerging norms. *Digital Journalism* DOI:10.1080/21670811.2014.906927.

Newman N, Fletcher R, Eddy K, Robertson CT & Nielsen RK (2023) Reuters Institute Digital News Report 2023. URL https://reutersinstitute.politics.ox.ac. uk/sites/default/files/2023-06/Digital_ News_Report_2023.pdf.

Nikolaev AG, Porpora D, Coffman N & Elliott-Maksymowicz K (2023) Hate Speech as a Form of Entertainment: An Unexpected Support for the Gratification Hypothesis on Twitter. *Atlantic Journal of Communication* DOI:10.1080/ 15456870.2023.2253344.

Niñoles G A & Ortega-Giménez C (2020) Discurso del odio en radio: Análisis de los editoriales de las cadenas COPE y SER tras la llegada del Aquarius a España. *Miguel Hernández Communication Journal* 11: 117–138. DOI:10.21134/mhcj. v11i0.317.

Noriega CA & Iribarren FJ (2013) Towards an empirical analysis of hate speech on commercial talk radio. *Harvard Journal of Hispanic Policy* 25: 69–96. URL https://link.gale.com/apps/doc/A412800595/AONE?u=univ&sid=bookmark-AONE&xid=5f707341

Olausson U (2017) The Reinvented Journalist: The discursive construction of professional identity on Twitter. Digital Journalism 5(1): 61–81. DOI:10.1080/21670811.2016.1146082.

Oliveira N, Costa J, Silva C & Ribeiro B (2020). Retweet Predictive Model for Predicting the Popularity of Tweets. In: A Madureira, A Abraham, N Gandhi, C Silva., M Antunes (eds) *Proceedings of the Tenth International Conference on Soft Computing and Pattern Recognition* (SoCPaR 2018). SoCPaR 2018. Advances in Intelligent Systems and Computing, 942. Springer, Cham. DOI:10.1007/978-3-030-17065-3_19

Orihuela JL (2015) Los medios después de Internet. Barcelona: UOC.

Palau-Sampio D & Carratalá A (2022) Injecting disinformation into public space: pseudo-media and reality-altering narratives. *Profesional de la información* / *Information Professional* 31(3). DOI:10.3145/epi.2022.may.12.

Pancer E & Poole M (2016) The popularity and virality of political social media: hashtags, mentions, and links predict likes and retweets of 2016 U.S. presidential nominees' tweets. *Social Influence* 11(4): 259–270. DOI:10.1080/15534510.2016.1265582

Paz MA, Montero-Díaz J & Moreno-Delgado A (2020) Hate speech: A Systematized Review. *SAGE Open* 10(4). DOI: 10.1177/2158244020973022.

Pérez JM, Furman DA, Alonso Alemany L & Luque FM (2022) RoBERTuito: a pre-trained language model for social media text in Spanish. In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, pp. 7235– 7243. URL https://aclanthology.org/2022. Lrec-1.785.

Peña-Fernández S, Larrondo-Ureta A. & Morales-i-Gras J. (2023). Feminism, gender identity and polarization in TikTok and Twitter. [Feminismo, identidad de género y polarización en TikTok y Twitter]. *Comunicar* 75, 49-60. DOI:10.3916/C75-2023-04

Piñeiro Otero T & Martínez-Rolán X (2021) Say it to my face: Analysing hate speech against women on Twitter. *Profesional de la información* / *Information Professional* 30(5): 1–17. DOI: 10.3145/epi.2021.sep.02.

Pla F, Hurtado LF, González JA, Ahuir V, Segarra E, Sanchis E, Castro MJ & García F (2022, September 21-23). GUAITA: Monitorización y análisis de redes sociales para la ayuda a la toma de decisiones [Conference session]. Proceedings of the Annual Conference of the Spanish Association for Natural Language Processing: Projects and Demonstrations (SEPLN-PD 2022), A Coruña, Spain. URL https://ceur-ws.org/Vol-3224/paper19.pdf

Richardson-Self L. (2018). Woman-hating: on misogyny, sexism, and hate speech. *Hypatia* 33(2): 256–272. DOI: 10.1111/hypa.12398.

Rodis PC (2021) Let's (re)tweet about racism and sexism: responses to cyber aggression toward Black and Asian women. Information, *Communication & Society* 24(14): 2153–2173. DOI:10.1080/1369118X.2021.1962948.

Rodríguez-Fernández L (2019) Desinformación: retos profesionales para el sector de la comunicación. El profesional de la información 28(3). DOI:10.3145/epi.2019.may.06.

Salvador M (2021) La propaganda, la desinformación y el populismo en las elecciones catalanas de 2021. Estudio de caso: la campaña de Vox en Twitter. Universitat Autònoma de Barcelona. URL https://ddd.uab.cat/record/249505

Sampaio-Dias S, Silveirinha MJ, Garcez B, Subtil F, Miranda J & Cerqueira C (2024) Journalists are Prepared for Critical Situations… but We are Not Prepared for This": Empirical and Structural Dimensions of Gendered Online Harassment. *Journalism Practice* 18(2): 301–318. DOI:10.1080/17512786.2023.2250755.

Sánchez-Meza M, Schlesier Corrales L, Visa Barbosa M & Carnicé-Mur M (2023) ¿De redes sociales a redes del odio? Análisis de la conversación digital en Twitter sobre la ministra de Igualdad española Irene Montero. *Estudios sobre el Mensaje Periodístico* 29(3): 717–736. DOI:10.5209/esmp.87271.

Shane T, Willaert T & Tuters M (2022) The rise of "gaslighting": debates about disinformation on Twitter and 4chan, and the possibility of a "good echo chamber". *Popular Communication* 20(3): 178–192. DOI:10.1080/15405702.2022.2044042.

Sheffer ML & Schultz B (2010) Paradigm Shift or Passing Fad? Twitter and Sports Journalism. *International Journal of Sport Communication* 3: 472–484. DOI:10.1123/ijsc.3.4.472.

Shifman L (2007) Humor in the Age of Digital Reproduction: Continuity and Change in Internet-Based Comic Texts. *International Journal*

of Communication 1: 187–209. URL https://ijoc.org/index.php/ijoc/article/view/11.

Slagle M (2009) An Ethical Exploration of Free Expression and the Problem of Hate Speech. *Journal of Mass Media Ethics* 24(4): 238–250. DOI:10.1080/08900520903320894.

Steinert J, Koch L & Pfeffer J (2023) Online Hate against Members of the European Parliament. Technical University of Munich. URL https://www.hfp.tum.de/fileadmin/w00cjd/globalhealth/_my_direct_uploads/EUP_OnlineHate__3_.pdf

Strömbäck J (2005) In Search of a Standard: four models of democracy and their normative implications for journalism. *Journalism Studies* 6(3): 331–345. DOI:10.1080/14616700500131950.

Suau-Gomila G, Pont-Sorribes C & Pedraza-Jiménez R (2020) Politicians or influencers? Twitter profiles of Pablo Iglesias and Albert Rivera in the Spanish general elections of 20-D and 26-J. Communication & Society 33(2): 209–225. DOI: 10.15581/003.33.2.209-225.

Tejedor S, Carniel-Bugs R & Giraldo-Luque S (2018) Los estudiantes de Comunicación en las redes sociales: estudio comparativo entre Brasil, Colombia y España. Transinformacão 30(2): 267–276. DOI:10.1590/2318-08892018000200010.

Tenenboim O & Kligler-Vilenchik N (2020). The meso news-space: Engaging the news between the public and private domains. *Digital Journalism* 8(5), 576-585. DOI:10.1080/21670811.2020.1745657.

Tortajada I, Comas D'Argemir i Cendra D & Martínez Corcuera R (2014) Inmigración, crisis económica y discursos radiofónicos: Hacia un lenguaje excluyente. *Estudios sobre el Mensaje Periodístico* 20(2): 899–916. DOI:10.5209/revESMP.2014.v20.n2.47063.

Urman A (2020) Context matters: Political polarization on Twitter from a comparative perspective. Media, Culture & Society 42(6): 857–879. DOI:10.1177/0163443719876541.

Vale PDA & Serra JN (2019) Comportamiento del lenguaje de líderes políticos venezolanos en el uso de Twitter. *Signo y Pensamiento* 38(74): 1–15. DOI:10.11144/Javeriana.syp38-74.cllp.

Valera-Ordaz L (2017) Comparing the democratic value of Facebook discussions across the profiles of Spanish political candidates during the 2011 General Election. *Revista Internacional de Sociología* 75(1): e052. DOI:10.3989/ris.2017.75.1.15.119.

Vis F (2013) Twitter as a reporting tool for breaking news: Journalists tweeting the 2011 UK riots. *Digital Journalism* 1(1): 27–47. DOI:10.1080/21670811.2012.741316.

Wahl-Jorgensen K (2001) Letters to the editor as a forum for public deliberation: Modes of publicity and democratic debate. *Critical Studies in Media Communication* 18(3): 303–320. DOI:10.1080/07393180128085.

Ward S (2018) *Disrupting Journalism Ethics: Radical Change on the Frontier of Digital Media.* London: Routledge.

Ward S (2020) Journalism Ethics. In: Wahl-Jorgensen K & Hanitzsch T (eds.) The Handbook of Journalism Studies, 2nd edition. London: Routledge, pp. 307–323.

Wilson AE, Parker VA, Feinberg, M (2020) Polarisation in the contemporary political and media landscape. *Current Opinion on Behavioral Sciences* 34: 223–228. https://doi.org/10.1016/j.cobeha.2020.07.005

Zhang X & Ho JCF (2022) Exploring the Fragmentation of the Representation of Data-Driven Journalism in the Twittersphere: A Network Analytics Approach. *Social Science Computer Review* 40(1): 42–60. DOI:10.1177/0894439320905522.