Check for updates

RESEARCH ARTICLE

Platform governance and public contestation in *X*: Media and user responses to *Meta*'s content moderation shift in Spain, the United States and Turkey

[version 1; peer review: 1 approved with reservations]

Ana Maria Zafra Arroyo[1], Pablo Pastora Estébanez[2]

[1]Journalism Department, University of Malaga, Málaga, Andalusia, Spain
[2]Journalism Department, University of Malaga, Málaga, Andalusia, Spain

v1 **First published:** 12 Sep 2026, **6**:350
https://doi.org/10.12688/openreseurope.23086.1
Latest published: 12 Sep 2026, **6**:350
https://doi.org/10.12688/openreseurope.23086.1

Abstract

Background

Scholarship on digital and platform governance distinguishes different regulatory models, including a deregulatory approach in the United States centred on free speech concerns, a multistakeholder model in the European Union, and patterns often described as digital authoritarianism in Turkey. In this context, this study examines media reactions on platform X to Meta's decision to replace fact-checking certified by the International Fact-Checking Network (IFCN) with a community-notes model and analyses how leading media outlets and users in the United States, Spain, and Turkey responded to news posts reporting this change in content moderation policy.

Methods

The methodology combines quantitative indicators of engagement, lexical analysis, qualitative coding of tone and themes, extraction and examination of user replies and continuous polarity modelling. The dataset includes posts from major news organisations and 1,052 user replies.

Results

The findings show cross-country variation in media framing and audience engagement. Media outlets in the United States and Spain produced posts combining neutral and critical tones and displayed

Open Peer Review

Approval Status ?

	1
version 1 12 Sep 2026	? view

1. **Mustafa AKBAYIR**, University of Akdeniz, Antalya, Turkey

Any reports and responses or comments on the article can be found at the end of the article.

diverse thematic patterns, while Turkish outlets focused mainly on descriptive reporting. User replies in the United States and Spain frequently introduced interpretative or critical elements, whereas replies in Turkey remained largely informational. Polarity modelling indicates that user responses are more strongly associated with the media outlet they address than with national context.

Conclusions

Media outlet framing emerges as a key mediating mechanism linking platform policy changes and audience responses across different governance contexts.

Keywords
audience response, computational social sciences, content moderation policy, digital governance, media discourse

This article is included in the COST Actions gateway.

This article is included in the RELINK for Democracy collection.

Corresponding author: Ana Maria Zafra Arroyo (amzafra@uma.es)

Author roles: Zafra Arroyo AM: Conceptualization, Investigation, Project Administration, Resources, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing; **Pastora Estébanez P**: Data Curation, Formal Analysis, Methodology, Software, Visualization

Competing interests: No competing interests were disclosed.

Grant information: This project has received funding from the European Union's Framework Programme for Research & Innovation as part of the COST Action [COST Action CA23114 (European Cooperation in Science and Technology) Ana María Zafra Arroyo], as supported by the COST Association (European Cooperation in Science and Technology). Ana María Zafra Arroyo is supported by an FPU predoctoral contract funded by the Spanish Ministry of Universities (grant FPU22-02894). Pablo Pastora Estébanez is supported by the research project Journalistic Applications of AI to Reduce Disinformation: Trends, Uses and Perceptions of Professionals and Audiences (PID2023-147486OB-I00), University of Málaga, Spain. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2026 Zafra Arroyo AM and Pastora Estébanez P. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Zafra Arroyo AM and Pastora Estébanez P. **Platform governance and public contestation in *X*: Media and user responses to *Meta*'s content moderation shift in Spain, the United States and Turkey [version 1; peer review: 1 approved with reservations]** Open Research Europe 2026, **6**:350 https://doi.org/10.12688/openreseurope.23086.1

First published: 12 Sep 2026, **6**:350 https://doi.org/10.12688/openreseurope.23086.1

1. Introduction

Donald Trump's return to the United States´(US) government has empowered high-tech companies in unprecedented ways. The most influential technology companies — known as GAFAM (*Google*, *Amazon*, *Facebook* [now *Meta*], *Apple* and *Microsoft*) — have acquired an unparalleled level of economic and political power. In 2020, these five companies' combined market capitalisation reached $5.35 trillion — exceeding the combined Gross Domestic Product of countries such as Japan, Germany, India and the United Kingdom (IG, 2021; Worldometer, 2023). These figures illustrate not only the magnitude of their dominance in the global market, but also the difficulties that governments face in regulating their activities.

In late 2025, the European Union imposed sanctions on the platform *X* for alleged non-compliance with its regulatory obligations under EU digital governance frameworks, a move that was followed by public statements from Donald Trump threatening the imposition of trade tariffs should the EU continue to challenge U.S.-based technology companies (Reuters, 2025). Against this backdrop, disputes over platforms governance intersect with geopolitical and economic pressures, fueling broader regulation tensions.

Announcements about content moderation policies have consequences that transcend the political and social spheres, as illustrated by Mark Zuckerberg's 2025 announcement changing Meta's content-moderation policy in US, replacing the IFCN-certified fact-checking system with a community-notes model. Around 60% of Internet users — approximately 3.2 million people — are active on at least one *Meta* platform, including *WhatsApp*, *Facebook*, *Instagram*, *Messenger* and *Threads* (Rogan, 2025). Therefore, this announcement is not just a business decision.

The United States is increasingly embracing a path toward deregulation. One of the arguments in favor of this shift is the assertion that content moderation practices enforced by verification agencies on social media platforms infringe upon citizens' fundamental right to freedom of expression. Critics contend that these agencies may function as instruments of censorship, selectively determining which voices and opinions are permissible in the public discourse (Adam, 2025). This issue becomes even more complex when considering the role of social media in various political systems.

In hybrid regimes or dictatorships, governments often manipulate social media channels to suppress dissenting views and control the narrative, effectively using these platforms as tools for censorship. However, it is important to note that, despite the prevalence of such practices in authoritarian contexts, there is a lack of empirical evidence suggesting that similar practices are being employed in democratic systems.

Indeed, entities like the European Union are worried about finding the right equilibrium between ensuring a secure online space and protecting individual freedoms. Efforts are underway to address both illegal content, which includes the promotion of terrorism, child exploitation, and threats to national security, as well as legal yet harmful content, such as disinformation, conspiracy theories, and polarization, all while upholding proportionality in relation to the defense of the right to freedom of expression (Kalliris, 2024).

This study analyses how leading media outlets in Spain, the US and Turkey reacted on platform *X* to *Meta*'s change in content-moderation policy, and examines how users responded to these posts.

1.1 Digital governance, platform governance, and public discourse in social media

Digital governance refers to the institutional design required to govern digital technologies, stressing governance mechanisms that manage risks and enhance public well-being (Filgueiras and Almeida, 2021). Models of digital governance range from centralised approaches to more decentralised or libertarian schemes, where the private sector shapes the digital ecosystem. Thus, there is a certain consensus on identifying at least three models of governance (O'Hara & Hall, 2018; Rone, 2021): the Chinese or authoritarian model; the Californian or Silicon Valley open internet model, and the European co-regulation or Brussels bourgeois model.

The Chinese model is characterised by pragmatism, state centralism, and a preventive policy. According to Cuihong and Liting, China is characterised by the implementation of strict regulations to mitigate social and cybersecurity risks, such as the enforcement of licences for online services, the use of automated filters and content censorship through the use of keywords. However, the internet is seen as a fundamental tool for economic development. The government exerts near-total control over the network, yet internet users, companies and social organisations work together under its directives to achieve more effective governance (2021).

In contrast, the Californian model gives companies enormous leeway to develop their business, even if this infringes individual rights such as data protection. This model is based on the idea of "move fast and break things — innovate until

the innovation is shown to be harmful" (O'Hara & Hall, 2018: 6). This model facilitates the formulation of flexible and dynamic policies, allowing for a rapid response to technological advances and market demands (Henton, 2001). The adoption of the "Internet freedom" policy by the US State Department as a soft power strategy involved the active participation of Silicon Valley companies, who promoted the principles of an open internet while aligning themselves with the government's interests in preserving US leadership in digital governance (Baykürt, 2022).

For its part, the European Union seeks to develop its own regulatory model based on multistakeholder governance and co-regulation, in which public institutions collaborate with relevant stakeholders in the design and implementation of policies that directly affect them. The EU does not fully trust market self-regulation, often considering it to be disruptive and a source of social costs. Although critics argue that this model suffers from excessive and sometimes contradictory regulations, which prevent it from being effective and competitive (Bauer, 2023), regulatory intervention is considered the way to achieve balanced agreements (O'Hara & Hall, 2018).

Within the broader framework of digital governance, the concept of platform governance has gained importance in the last decade. According to Luna-Reyes: "The first set of applications that challenge traditional governance mechanisms and information flows in society are social media private platfoms (2017: 199). Gorwa (2019) characterizes the platform governance concept as the layered governance relationships that link platform practices, policies, affordances, and external political forces. This interaction produces asymmetric power relations, particularly when large platforms operate transnationally across diverse political regimes, adapting their governance practices to different regulatory and institutional contexts. "Social media platforms are also private companies that are actively looking for ways to monetize user activity, mostly through marketing. In this way, all public conversations on these platforms are mediated by algorithms and rules designed to maximize the company's profit rather than benefit the society" (Luna-Reyes, 2017: 200).

Concerns such as disinformation, content moderation, algorithmic transparency, and the protection of freedom of expression are central to debates on platform regulation due to their impact on society. Bradshaw and Howard (2019) describe how organised manipulation on social media has shaped a "global order of disinformation", in which state and non-state actors deploy large-scale computational propaganda strategies. These findings provide a basis for understanding how governments and political parties use armies of automated accounts and trolls for electoral purposes and to control public discourse. According to Badillo et al.: "Although it may seem paradoxical, political systems with greater freedom of expression are also more vulnerable than autocracies or dictatorships to disinformation campaigns because they benefit from the free circulation of information, one of the cornerstones of democratic systems" (2023: 54).

Phillips and Milner (2021) provide a framework for interpreting polarisation and conspiracy theories, demonstrating how digital environments can reinforce the dynamics of otherness and discursive radicalisation. In the Turkish context, Saka (2018) describes the activities of 'AKtrolls', revealing how fake identities and automated profiles are systematically employed by government-aligned individuals to influence public discourse. These contributions are particularly relevant to the study of public discourse on platforms such as X (formerly Twitter), as they enable the issue of content moderation to be addressed as one of the main features of digital public spaces, insofar as it defines the boundaries of truthfulness and the circulation of extremist content, as well as the scope for expression in platform-mediated environments.

1.2 Content moderation on social media platforms

According to Gomes and Sultan, "content moderation is a key quality regulation of online spaces; it is the main mechanism through which social media platforms examine content against their terms of services and community guidelines to determine its suitability" (2024: 2). Social media platforms typically use hybrid moderation systems that combine Artificial Intelligence (AI) and human oversight to manage content. These systems use technologies such as Natural Language Processing (NLP) and Convolutional Neural Networks (CNN) to detect harmful content, such as hate speech and misinformation (Nehe et al., 2024; Govindankutty & Kumar, 2024).

As technologies advance, Large Language Models (LLMs) are being implemented that can learn to identify harmful content more effectively than traditional methods from very few training examples (Bonagiri et al., 2025). However, these systems suffer from problems related to the quality of training datasets and bias (Car, 2025). This is why platforms have incorporated IFCN-certified fact-checkers into their content moderation departments[1].

[1]The International Fact-Checking Network (IFCN) is a global initiative hosted by the Poynter Institute that sets professional standards for fact-checking organizations. It operates a rigorous certification program that evaluates transparency, methodological soundness, nonpartisanship, and accountability.

Recently, there has been a trend among digital platforms to view content moderation as a potential threat to freedom of expression (Gosztonyi et al., 2025) resulting in their deregulation. In the United States, *Meta* and *X* have replaced the use of independent fact-checkers with community notes system, which allow users to comment on the accuracy of posts. However, studies have shown that relaxing *X*'s moderation policies has increased hate speech and the formation of polarised communities (Arun et al., 2024).

In non-democratic countries, governments control the content circulating on networks using measures such as word filters. But, on the other hand, by delegating this control to platforms, "there are no guarantees that these providers will be motivated by the benevolence" (Kalliris, 2024: 413). Sánchez Muñoz therefore proposes various measures to protect freedom of expression in content moderation. These include clearly defining the harms to be prevented based on evidence, ensuring transparency in the process to enable assessment of its impact, guaranteeing the right of appeal for authors of deleted content, and mitigating the risks of excessive automation through human supervision (Sánchez Muñoz, 2020: 120). A study by the University of Oxford and the Technical University of Munich found that most respondents would prefer a social media platform that prevents the spread of hate and disinformation, even if it means sacrificing absolute freedom of expression (Zilinsky et al., 2025).

In discussions surrounding the regulatory landscape of social media, authors like Kohl (2024) have raised concerns regarding the immunity that social media platforms hold within the European Union. This immunity is troubling given their sophisticated algorithmic recommendation systems. Rather than functioning as mere conduits for information, these platforms are proactive architects of the digital environment, influencing what users see, engage with, and ultimately believe (Gillespie, 2018). The business models underpinning these social media giants are intricately tied to the proliferation of disinformation, echo chambers, and societal polarization (Bakshy et al., 2015; Lazer et al., 2018; Pariser, 2012). By prioritizing user engagement and maximizing profits, social media platforms employ algorithms that curate personalized content based on individual users' preferences. This targeted approach not only keeps users engaged for longer periods but also reinforces existing biases.

1.3 Media system in Spain, Turkey and US

Numerous studies have demonstrated the relationship between a country's political and media systems. For example, it has been shown that in countries with higher levels of democracy, the media operate with greater freedom (Herman & Chomsky, 1988; Hallin et al., 2008). Spain, US, and Turkey represent three countries with different political systems and media landscapes. While the former two are considered democratic states, Turkey is not recognised as a fully free country due to its government's authoritarian drift in recent years (Freedom House, 2024).

The Eurasian country is classified as a flawed democracy or hybrid regime. Around 90% of traditional media outlets, including newspapers, radio and television, support the government's ideology (Reporters Without Borders, 2024). "The government is similarly taking advantage of the potential of social media, employing a dual strategy of communication and mobilisation. (…) A twofold strategy aimed at circulating fictional or misleading political narratives and systematically harassing the regime's opponents"(De Cosentino & Alikasifoglu, 2019: 93). Although digital media outlets have more diverse editorial lines, access to press accreditation is controlled by a government-dependent body that issues press cards (BBC News Türkçe, 2022). Consequently, digital media journalists may face restrictions on their accreditation if their publications are overly critical.

Spain, for its part, falls within what Hallin and Mancini (2008) call the Polarised Pluralist System: a media model characterised by low newspaper circulation, limited professionalisation of journalism, strong political alignments, and marked state interventionism. This manifests in the regulation of media structure and ownership, as well as the control and management of public media (Hallin & Mancini, 2008; Seijas Costa et al., 2024; Serrano Puche, 2017).

The 2008 debt crisis had a significant impact on Spanish media conglomerates, resulting in widespread job losses, reduced staff numbers, and deteriorating working conditions for professionals (García Madariaga & Arasanz Esteban, 2019). This crisis-stricken context led to structural changes in the industry, encouraging the adoption of digitalisation and new forms of information production (García-Santamaría & Pérez-Serrano, 2020). Polarisation is another distinctive feature of the Spanish media system. Ideological alignment between the media and political parties means that the tension of political debate is transferred to the media.

The relaxation of moderation policies on social media platforms like *X* has prompted Spanish news outlets, such as *La Vanguardia*, to depart from this social network. Despite being selected as a source of information by 16% of the population (Newman et al., 2024), *X* has become a platform rife with polarisation, disinformation, and the spread of hate speech (Majó-Vázquez & González-Bailón, 2022; Reuters, 2024).

Finally, the United States is classified within the Liberal or North Atlantic model. The US media system is characterised by a high degree of liberalism and market-driven ownership. This approach has its roots in the First Amendment, which implicitly protects freedom of expression and limits government intervention in the media's operations.

However, the main criticism of this model is that media plurality and diversity are not guaranteed, as libertarianism prioritises corporate rights over public interests (Pickard, 2015). The Federal Communications Commission (FCC) is a regulatory body whose objectives include ensuring diversity of channels and opinions. In addition to facing pressure from an industry that favours deregulation, this body has had to contend with several court rulings that have hindered its work (Candeub, 2008; Fukuyama & Grotto, 2020).

While the EU has implemented comprehensive digital regulations such as the General Data Protection Regulation (GDPR) to address issues such as privacy and disinformation, the US has not adopted such an interventionist approach (Fukuyama & Grotto, 2020). Figures such as Donald Trump have exploited these platforms to circumvent traditional media and communicate directly with the public, altering the dissemination and reception of political messages (Harbi, 2024).

This evolution has raised questions about the credibility and reliability of information sources and has contributed to the fragmentation of public discourse (Dickerson, 2019). Furthermore, information is often approached from a sensationalist and entertainment-focused perspective, as evidenced by the coverage of political campaigns such as Donald Trump's in 2016 and 2024. By portraying his campaign as a spectacle rather than focusing on political issues, the media allowed him to dominate the narrative and attract his supporters (Harbi, 2024).

2. Objectives and hypotheses

The overall objective of this study is to examine media reactions on *X* to *Meta*'s shift from IFCN-certified fact-checking to a community-notes model, and to analyse how users respond to these posts in three different digital governance contexts (Spain, the US and Turkey).

To achieve this, the following Specific Objectives (SO) have been defined:

SO1: Identify and compare the main discursive tones and stance adopted by the media outlets in each country in their *X* posts reacting to *Meta*'s announcement.

SO2: Classify the predominant thematic and lexical–semantic patterns present in the media posts, highlighting cross-country differences.

SO3: To examine how user replies relate to the discursive framing of media posts across countries.

SO4. Examine how both media discourse and user-reply patterns relate to the broader regulatory frameworks of digital governance in the United States, Spain and Turkey.

Similarly, the following hypotheses (H) are proposed:

H1. In Turkey, where government control over digital platforms is highest, media outlets will adopt a more favourable stance toward *Meta*'s decision to relax content moderation oversight, and user replies will show comparatively lower levels of explicit disagreement or polarization.

H2. In Spain, where digital governance emphasises user rights and co-regulation, media outlets will express a more critical stance toward *Meta*'s announcement, and user replies will tend to reproduce or amplify this critical orientation.

H3. In the United States, media reactions will reflect ideological alignment: outlets aligned with Democratic positions will be more critical, while those closer to Republican positions will show neutrality or support; user replies will mirror this partisan divide, exhibiting higher levels of polarity.

H4. Higher polarity scores and stronger lexical–semantic differentiation will emerge in media systems characterised by greater ideological fragmentation (US), whereas more centralised or government-aligned systems (Turkey) will display lower polarity and more homogeneous semantic patterns.

3. Methodology

This research has developed a methodology that combines automated data extraction techniques, descriptive quantitative analysis, semantic analysis and qualitative content analysis. This mixed approach is supported by computational tools that allow both the discourse emitted by the media and the reactions and communicative practices of users on *X* (formerly *Twitter*) to be examined. This approach falls within the field of computational social science, which is defined as the combination of computational methods and theoretical frameworks from the social sciences to systematically study complex phenomena (Conte et al., 2012; Edelmann et al., 2020). While the volume of data is not massive, scraping techniques, automatic language processing and computational analysis of digital interactions enable a replicable approach to the subject.

The methodological process can be divided into five main phases:

(1) Selection of the sample and countries under study.

(2) Automated collection of news headlines covering *Meta*'s announcement as shared within *X* posts, using web-scraping techniques.

(3) Data processing, cleaning, and quantitative analysis, including dataset structuring and statistical metrics.

(4) Software-assisted qualitative analysis using *Atlas.ti* to manually code key posts from the five largest media outlets in each country, followed by co-occurrence matrix generation and heatmap-based comparative visualization.

(5) Extraction and analytical evaluation of user replies linked to the collected X posts.

(6) Lexical–semantic analysis and polarity modeling, incorporating sentiment detection and the construction of tone or polarization models.

3.1 Data selection and sample design

X was selected as the platform for analysis due because it is the social media site most used by journalists for professional purposes according to a Pew Research Center survey (Jurkowitz & Gottfried, 2022). Additionally, *X*'s content moderation model is similar to *Meta*'s proposed approach, which relies on user participation through community notes rather than exclusively on external fact-checkers.

In terms of countries, Spain, the United States, and Turkey were chosen as they represent three different models of digital governance and levels of press freedom, as well as having divergent media ecosystems. This allows for a comparative analysis of narratives and citizen replies to the change in *Meta*'s content moderation policies.

For each country, media outlets with the greatest digital reach were selected according to the 2024 Digital News Report, provided they had an official *X* account. The final sample included the following media outlets from different countries: five Spanish outlets (*El País*, *El Mundo*, *20minutos*, *eldiario.es* and *OkDiario*), five American outlets (*The New York Times*, *The Washington Post*, *USA Today*, *The Wall Street Journal* and *New York Post*) and five Turkish outlets (*Sözcü*, *Habertürk*, *Haberler.com*, *Ensonhaber* and *SonDakika*).

3.2 Extraction and cleaning of news headlines as shared within X posts

The data was extracted using the *TweetScraperR (R)* package via web scraping, with advanced searches encoded in URLs that replicate manual queries on the platform. All tweets posted by the selected official accounts since 6 January 2025 that included references to "Meta", "Facebook", "Instagram" or "Mark Zuckerberg", as well as terms related to content moderation: "censorship", "verification", "algorithmic", "fact-checking" were retrieved.

The obtained data underwent a cleaning process:

- removal of irrelevant columns.
- format standardisation.

- identification and removal of duplicates.
- classification by country and media.
- extraction of interaction metrics (likes, reposts and replies).

Finally, 56 headlines were obtained to form the basis of the initial descriptive and interpretative analysis.

3.3 Quantitative analysis of news headlines as shared within X posts

The statistical analysis was carried out entirely in *R* using the *dplyr*, *tidyr*, *ggplot2* and *kableExtra* packages. Descriptive metrics were calculated for each country and medium, including:

- total number of publications.
- average number of replies.
- average number of reposts.
- average number of likes.
- engagement distribution.

A lexical analysis was also performed using *TidyText* and *Text2Vec*. The text underwent pre-processing (normalisation, removal of stop words, lemmatisation and translation where necessary) in order to extract bigrams and analyse patterns of semantic co-occurrence. The results enabled the dominant narratives in each country to be characterised, as well as the centrality of concepts such as “moderation”, “censorship”, “fact-checking” and references to key political figures.

3.4 Qualitative analysis of news headlines as shared within X posts

In addition to the computational approach outlined above, a phase of software-assisted qualitative analysis was conducted to identify and compare the main discursive tones and stances adopted by media outlets in each country in their posts on *X* reacting to *Meta*’s announcement. This stage adopted an inductive-interpretative logic, centring on the manual coding of the most pertinent tweets published by the five media outlets with the widest reach in each country.

Atlas.ti was used to organise and analyse the qualitative data by tone, discursive themes, and analytical codes. Co-occurrence matrices crossing countries with tone and thematic categories were generated and exported to Microsoft Excel, where they were visualised using heatmaps. To support cross-national comparison of codes, a comparative table was produced using normalised percentages (relative to each country’s total coded segments), and the most frequent non-tone codes were reported in the text. This qualitative approach captured contextual nuances and specific narratives that automated semantic analysis may overlook, adding an interpretative dimension to understanding media positioning in relation to the predominant digital governance models in each country.

3.5 Collection of replies

To analyse the public’s reception of the headlines, all replies associated with the collected posts were extracted. This process was carried out using the TweetScraperR (R) package, which provides stable access to *X*’s unofficial API via multiple accounts authenticated using cookies and session parameters.

The steps were as follows:

1. Extraction of the tweet identifier (tweet_id) from the URL of each publication.
2. Use of the tweet_replies() function to retrieve all replies associated with each tweet identifier.
3. Converting the data to dataframe format using to_dataframe() and consolidating it into a single dataset.
4. Integrating contextual variables from the original tweet (media, country, date and headline) into each reply.
5. Cleaning the dataset by removing empty replies, duplicates and replies without valid linguistic content.

The final dataset includes several hundred replies, which is sufficient for robust quantitative, semantic and comparative analysis.

3.6 Sentiment analysis of user replies

To examine the affective orientation of user reactions to the changes in *Meta*'s content moderation policy, a computational sentiment analysis was applied to user replies posted on *X* in response to media coverage of the announcement.

Given the multilingual nature of the corpus—comprising replies written primarily in Spanish and English, as well as a smaller number of messages in other languages—all user replies were automatically translated into English prior to sentiment analysis. This step was implemented to ensure linguistic standardization and comparability across national contexts, and to allow the application of a single high-performing sentiment classification model.

Translation was performed using a neural machine translation model based on the No Language Left Behind architecture (NLLB-200, distilled version), specifically facebook/nllb-200-distilled-600 M. This model is designed for multilingual translation and has demonstrated robust performance across a wide range of language pairs. In the present study, translation is conceptualized as a normalization procedure, aimed at preserving the general semantic content of user replies while facilitating downstream sentiment analysis.

While automatic translation may entail a partial loss of pragmatic or cultural nuance, its use is methodologically justified in comparative computational research, where consistency of linguistic input is a prerequisite for valid cross-country comparisons.

Following translation, sentiment analysis was conducted using a transformer-based model derived from RoBERTa, fine-tuned specifically for sentiment classification on social media data. The model employed was cardiffnlp/twitter-roberta-base-sentiment-latest, which has been trained on large-scale Twitter corpora and optimized to capture the linguistic characteristics of short, informal, and opinionated texts.

The model assigns each reply to one of three sentiment categories—positive, negative, or neutral—and outputs the corresponding posterior probabilities for each class. This probabilistic output enables a more nuanced analysis than categorical classification alone and supports the construction of continuous sentiment indicators.

To capture the intensity and direction of affective orientation in a more granular manner, the probabilistic outputs of the model were transformed into a continuous polarity index, defined as the difference between the predicted probability of positive sentiment and the predicted probability of negative sentiment:

$$\text{Polarity} = \text{P}(\text{positive}) - \text{P}(\text{negative})$$

This index ranges from -1 (strongly negative orientation) to $+1$ (strongly positive orientation), with values close to zero indicating neutral or ambivalent sentiment. Polarity scores were computed at the level of individual replies and subsequently aggregated by country and by media outlets using robust summary statistics (median and interquartile range), in order to reduce the influence of extreme values.

The quantitative sentiment analysis was applied exclusively to replies associated with media outlets from Spain and the United States. In the case of Turkey, the number of available user replies was insufficient to support statistically comparable sentiment estimates. Consequently, Turkish replies were excluded from the quantitative sentiment analysis and were examined solely through descriptive and qualitative analyses. This methodological decision reflects a commitment to analytical rigor, ensuring that each technique is applied only where the empirical structure of the data allows for reliable interpretation.

4. Analysis and results

This section presents the results derived from the computational and qualitative procedures described in the methodological framework. Each subsection corresponds to one of the study's specific objectives. The analyses combine descriptive statistics; lexical and co-occurrence modelling; qualitative coding assisted by Atlas.ti; extraction and examination of user replies, and polarity modelling using RoBERTa.

A preliminary descriptive analysis was conducted to characterise the dataset before proceeding to the objective-specific analyses. Interaction metrics (replies, reposts and likes) were calculated using Python after scraping the posts through the

academic API of platform *X*. Mean engagement levels were computed for the full sample: 59.49 replies, 77.61 retweets and 190.05 likes per post. Posting frequency was also extracted from the dataset. US outlets published the highest number of posts (*New York Post* = 14; *Washington Pos*t = 12), followed by *El País* and *eldiario.es* (8 each), and *The Wall Street Journal* (5). Turkish outlets showed lower posting frequency.

These descriptive indicators provide the baseline distribution from which the subsequent lexical, thematic and discursive analyses were computed.

4.1 Discursive tone and stance of media outlets

The tone analysis was conducted manually using an inductive coding scheme in Atlas.ti. Each post was coded according to the explicit stance expressed towards Meta's content-moderation policy change. Posts were classified as critical or neutral. No supportive posts were identified in the corpus.

Figure 1 shows cross-national differences. In the United States, tone was evenly split between critical (n = 10) and neutral posts (n = 10). In Spain, critical posts (n = 7) were slightly fewer than neutral posts (n = 8). In Turkey, no critical posts were observed (n = 0), and all posts were coded as neutral (n = 5).

These results derive from qualitative coding focused on evaluative markers (e.g., explicit criticism of *Meta*'s decision, references to censorship, …). Beyond tone, the analysis examined the distribution of discursive themes across countries. As shown in Figure 2, themes related to political context and criticism of institutional or corporate actors were most prominent in US media coverage, moderately present in Spanish outlets, and largely absent from Turkish media, which displayed a more informative and less critical perspective.

Table 1 compares the relative prevalence of the most frequent non-tone codes across countries. In the United States, posts more often referenced changes to verification policies (16.7%) and mentioned Zuckerberg (16.7%), whereas in Spain the most salient codes within this subset were mentions of Trump (8.1%) and comparisons with *X* (8.1%). In Turkey, references to changes to verification policies were proportionally more frequent (23.1%) than in the US and Spain, and

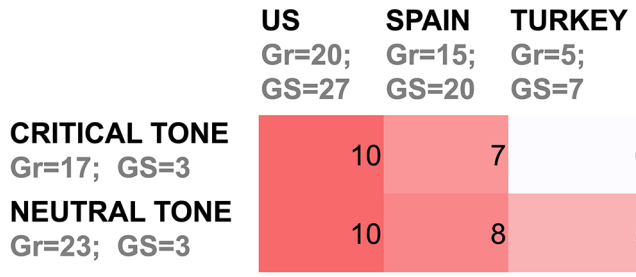


Figure 1. Heatmap of discursive tone by country.

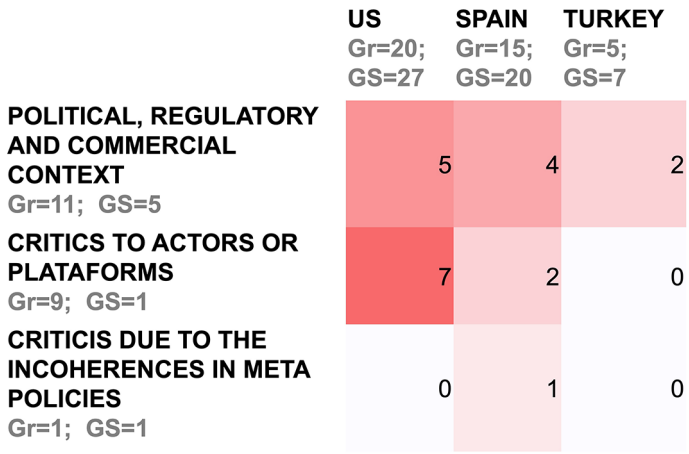


Figure 2. Heatmap of discursive themes by country.

Table 1. Comparative distribution of discursive codes across countries.

Código	US (%)	Spain (%)	Turkey (%)
Changes to verification policies	16.7	8.1	23.1
Mention of Zuckerberg	16.7	0	15.4
Criticism of Zuckerberg/Meta	11.7	5.4	0.0
Mention of Trump	6.7	8.1	0.0
Freedom of expression	6.7	2.7	7.7
Permissiveness towards hate speech, violence, etc.	3.3	5.4	7.7
Comparison with X	1.7	8.1	0

Source: Author's own analysis using Atlas.ti.
Note: Values represent percentages of coded posts within each country (US n = 60; Spain n = 37; Turkey n = 13). Only the most frequent non-tone codes are reported.

mentions of Zuckerberg also represented a notable share (15.4%). Criticism of Zuckerberg/*Meta* was present in the US (11.7%) and Spain (5.4%) but not observed in Turkey within this subset.

4.2 Thematic and lexical–semantic patterns in media posts

Lexical frequencies, bigrams and co-occurrences were computed using Python's NLTK and spaCy libraries. After cleaning and tokenising the corpus, frequency lists and bigram models were generated to detect dominant terms and associations. Co-occurrence matrices were computed using sliding-window methods and visualised with heatmaps. In the Turkish case, lexical analysis showed frequent use of *meta*, *zuckerberg*, *facebook*, *mark*, *content* and *instagram.* Bigrams (*mark zuckerberg*, *facebook instagram*, *fact checking*) were generated through Python.

In Turkey, the most frequent terms in the corpus were *meta*, *zuckerberg*, *facebook*, *mark*, *content* and *instagram.* Dominant bigrams—such as *mark zuckerberg*, *facebook instagram* and *fact checking*—were identified. In Spain, frequent terms included *meta*, *verification*, *facebook* and *content.* Co-occurrence patterns such as *facebook instagram* and *content moderation*, obtained through a word-window co-occurrence model. Additional bigrams, including *porn ads* and *trump threatens*, emerged through frequency-weighted bigram extraction.

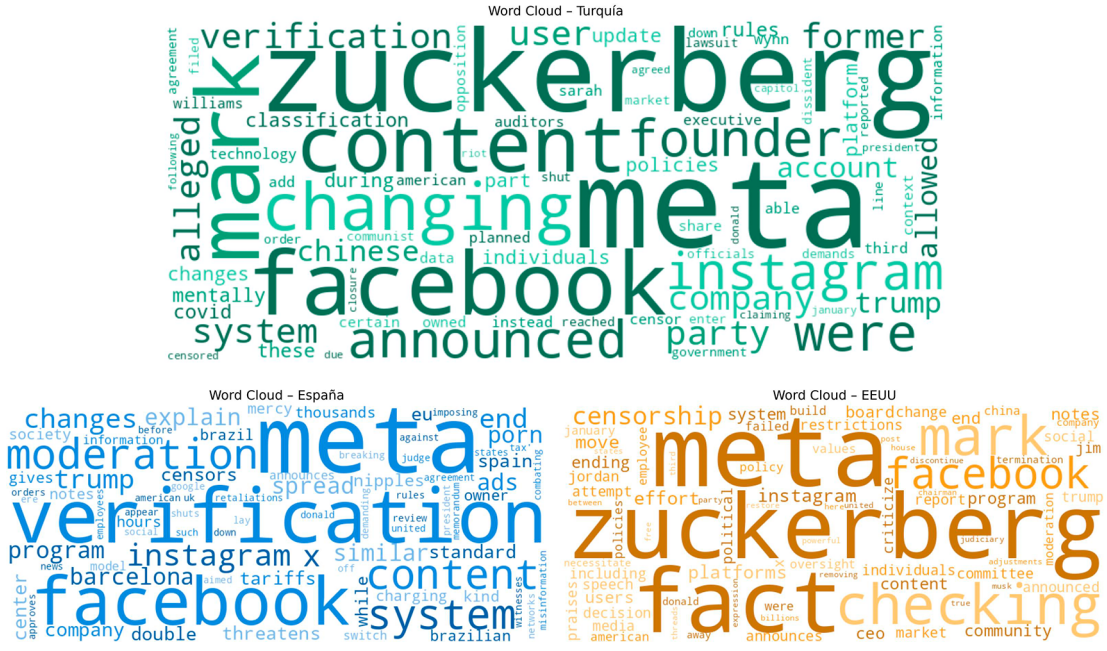


Figure 3. Word clouds of the most frequent terms by country.

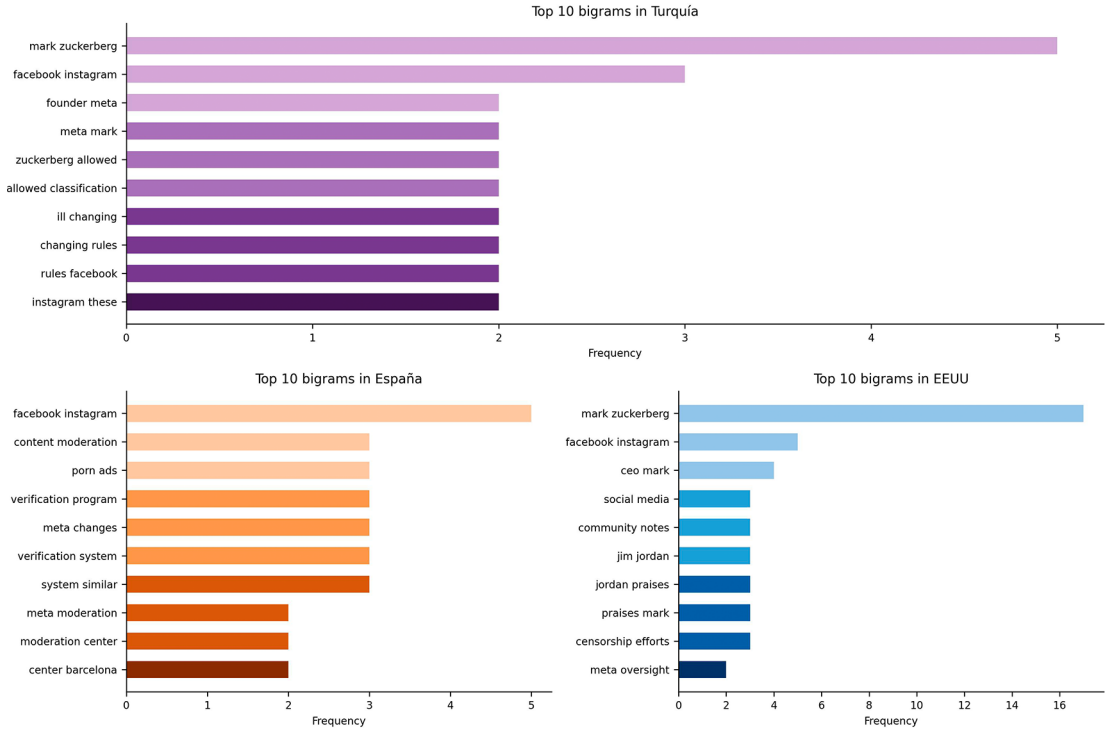


Figure 4. Top 10 bigrams by country.

As shown in Figure 3, the most frequent lexical items in the United States were *fact*, *meta* and *zuckerberg*. Bigram scoring in Python identified *fact checking* and *Mark Zuckerberg* as highly recurrent combinations (see Figure 4).

4.3 Analysis of user replies and audience engagement

User replies were extracted by scraping the reply threads associated with each media post, resulting in a dataset of 1,052 responses. This dataset was cleaned and tokenised using Python, and the replies were subsequently coded inductively in Atlas.ti to identify tone and discursive frames. Tone classification followed the same criteria applied to the media posts. The results show marked cross-national differences in how audiences responded to Meta's policy change. In the United States, replies frequently contained evaluative and critical markers, with users referencing "censorship," "bias" or "political manipulation". For instance, in response to a *New York Post* headline announcing the end of fact-checking, one

Table 2. Descriptive statistics of user reply polarity (overall and by country).

Group	N	Mean polarity	SD	Median
Overall	1,046	−0.47	0.29	−0.52
Spain	290	−0.52	0.26	−0.52
United States	756	−0.56	0.25	−0.56

Source: Author's own elaboration using RoBERTa.

Table 3. Statistical tests on user reply polarity.

Comparison	Test	Statistic	df	p-value	Result
Reply polarity × Country	Mann–Whitney U	—	—	> .05	Not significant
Reply polarity × Headline polarity	Pearson r	−0.01	—	.91	Not significant
Reply polarity × Media outlet	χ^2	77.31	24	< .001	Significant

Source: Author's own elaboration using RoBERTa.

Table 4. Mean polarity of user replies by media outlet (N ≥ 5).

Media outlet	Country	N	Mean polarity	SD
@wsj	United States	57	−0.11	0.24
@elmundoes	Spain	24	−0.18	0.30
@nytimes	United States	36	−0.36	0.28
@el_pais	Spain	237	−0.40	0.18
@eldiarioes	Spain	27	−0.42	0.31
@nypost	United States	434	−0.48	0.14
@washingtonpost	United States	229	−0.50	0.13

Source: Author's own elaboration using RoBERTa

user reacted with a brief but expressive "whattt," while others expanded the critique by asserting that Zuckerberg "helped steal the 2020 election" or that he was attempting to "get in good with the Trump Administration."

In Spain, replies also displayed a predominantly critical orientation, combining irony, evaluative commentary and scepticism towards *Meta*'s governance decisions. For example, reacting to a *20 Minutos* post on *Meta*'s decision to drop moderation policies, one user asked whether Zuckerberg intended to "liberalise female nipples or only the disgusting far-right stuff," reframing content moderation as an issue of selective permissiveness. Another reply linked Meta's actions to political dynamics involving Elon Musk and Donald Trump, suggesting that fears of platform sanctions influenced the decision. Similarly, a response to an *El País* post about Trump's claims of censorship stated that combating disinformation merely served to "censor dissenters," thereby reframing verification as a mechanism of ideological control. These examples demonstrate that Spanish users frequently expanded the media frame by introducing political critique and social commentary.

User replies in Turkey largely mirrored the informational and descriptive tone of media posts, showing limited engagement with the governance or censorship implications of Meta's policy change. Most comments reiterated the information provided rather than introducing interpretative debate. Occasional reactions diverted discussion toward moral or hostile positions regarding actors mentioned in the news, but these did not constitute the dominant response pattern.

4.4 Media ecosystems and interpretative communities in shaping user sentiment

Sentiment polarity was computed after translating all replies into English and applying a transformer-based sentiment classifier. Quantitative analyses focus on Spain and the United States, as the number of replies associated with Turkish

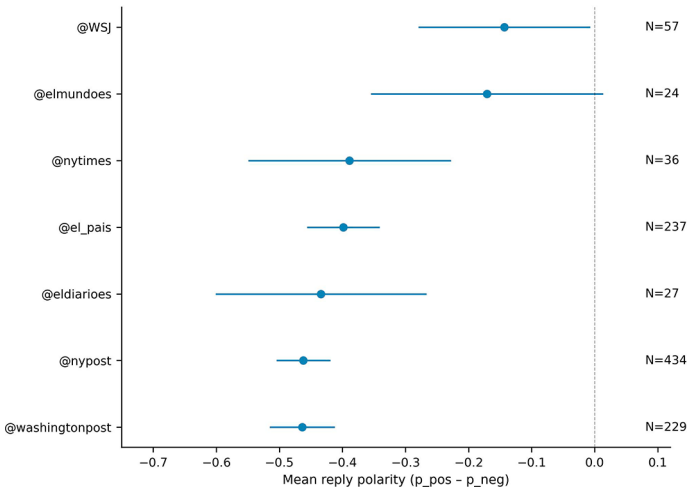


Figure 5. Mean polarity of user replies by media outlet.

media outlets was insufficient for robust statistical modelling and is therefore addressed exclusively in the qualitative analysis.

At the aggregate level, user replies display a predominantly negative emotional orientation. As shown in Table 2, the global mean polarity is clearly below the neutral threshold, with a negative median indicating that critical reactions dominate the conversation. The dispersion of values suggests a heterogeneous emotional response, yet consistently skewed towards negativity. These results confirm that *Meta*'s policy change elicited a largely critical reaction among users across the dataset as a whole.

When disaggregated by country, both Spain and the United States exhibit very similar patterns of reply polarity. As reported in Table 2, average polarity values in both countries remain strongly negative, with only minor differences in central tendency and dispersion. Formal statistical tests indicate that these differences do not reach statistical significance, suggesting that the affective tone of user reactions does not vary substantially between the two national contexts. Despite differences in media systems and regulatory traditions, the emotional response to *Meta*'s content moderation changes appears broadly comparable in Spain and the United States. This finding indicates that national context alone is insufficient to explain variation in user sentiment and points to the relevance of other structural factors in shaping audience reactions.

Additional analyses examined whether the emotional tone of media headlines influenced the polarity of user replies. As reported in Table 3 correlation analyses and regression-based robustness checks consistently show no statistically significant relationship between headline polarity and the polarity of replies. This result suggests that more negative or more positive headlines do not systematically generate correspondingly negative or positive audience reactions. Instead, user sentiment appears to be driven by factors beyond the immediate emotional framing of individual posts. The absence of a headline-level effect reinforces the need to consider broader contextual dynamics when analysing audience responses to platform governance debates.

In contrast to the lack of significant differences by country or headline polarity, clear and statistically significant variation emerges at the media outlet level. Descriptive statistics by outlet, reported in Table 4, reveal substantial differences in the mean polarity of user replies. While all outlets receive predominantly negative responses, the intensity of negativity varies markedly across media, ranging from near-neutral averages to strongly negative patterns.

A chi-square test of independence confirms a significant association between reply polarity category and media outlet ($\chi^2 = 77.31$, df = 24, $p < .001$), indicating that user reactions are not randomly distributed across outlets. Instead, specific media accounts systematically attract more negative or more moderate responses. These differences are visualised in Figure 5, which plots the mean polarity of user replies for each media outlet. Figure 5 shows a clear clustering of outlets according to the intensity of negative sentiment, highlighting the role of outlet-specific dynamics in structuring audience reactions.

Together, these results suggest that media ecosystems and audience alignments play a more decisive role in shaping user sentiment than national context or headline tone. User reactions appear embedded in media-specific interpretative communities, where prior orientations towards particular outlets influence how platform governance decisions are received and evaluated.

5. Discusion

In this article, aspects such as the level of media freedom in each country and the way in which each country approaches platform governance have been considered in order to assess whether differences exist between them. In this regard, although there are widespread concerns about freedom of expression in these contexts, platform companies frequently adopt a pragmatic stance in their operations within non-democratic environments, accepting governmental authority and complying with restrictive legal and political demands in order to maintain their market access. In Turkey, amendments to the national internet legislation require social media platforms to appoint a local legal representative within the country. This representative is obliged to cooperate with public authorities in the enforcement of domestic regulations, including content takedown requests and data-related obligations (Information Technology & Innovation Foundation, 2025).

Such regulations, widely criticized for their restrictive impact on freedom of expression, facilitates platform cooperation with state censorship mechanisms. According to the last Freedom House report (2025) about Turkey, social media companies, including *X*, suspended or restricted accounts belonging to independent media organizations and opposition politicians in response to government demands.

As Gorwa points out, it should not be overlooked that these network sites "are influential political actors with an obvious interest in preserving their dominant market positions" (Gorwa, 2019: 865). Less obvious, however, is the extent to which large digital platforms exert influence in democracies. Users may be free to think, speak, and participate, yet cases like the Facebook–Cambridge Analytica scandal illustrate how challenge democratic political processes can be challenged.

Moreover, the political influence of these high techs has generated an institutional pushback. Recently, leaders of the European Union publicly defended the autonomy of its digital regulation after President Donald Trump threatened retaliatory measures against the EU in response to enforcement actions under the Digital Services Act (Le Monde, 2025).

In this context, the new dilemma for the European Union is not limited to regulating its digital environment, but to position itself as an influential geopolitical actor in a global order increasingly shaped by non-European platform and artificial intelligence corporations. Control over large-scale data flows, algorithmic systems, and digital infrastructures has become an axis of contemporary power, one that will ultimately shape the capacity of democratic systems to govern themselves in the digital age.

6. Conclusion

This study examined how leading media outlets in the United States, Spain and Turkey framed *Meta*'s decision to replace IFCN-certified fact-checking with a community-notes model on platform *X*, and how users responded to these media posts. The analysis integrated quantitative metrics, lexical–semantic patterns, tone, thematic coding, user replies and polarity modelling.

The findings addressing SO1 reveal cross-national variation in media tone and stance. US and Spanish outlets combined neutral reporting with varying degrees of critical and politicised framing, whereas Turkish media maintained a consistently neutral and informational tone, with minimal engagement in evaluative or critical discourse. The absence of critical tone in the Turkish sample are consistent with broader findings on constrained information environments (Ertuna & Aşık, 2025; Yücel, 2025).

Regarding SO2, the lexical–semantic and thematic coding analyses also revealed variations across countries. In the United States, dominant lexical items and recurrent bigrams foregrounded issues of fact-checking and censorship, indicating a more conflict-oriented framing of Meta's policy change. Spanish media discourse, by contrast, was characterised by a regulatory vocabulary centred on verification, content moderation and institutional processes, reflecting a more governance-focused emphasis. In Turkey, lexical patterns were largely actor- and platform-centred, with frequent references to *Meta*, Zuckerberg and affiliated platforms, and a relative absence of evaluative or critical semantic associations.

With respect to SO3, in the United States and Spain, audience responses frequently challenged or expanded the discursive frames established by media outlets, introducing additional political interpretations, normative judgments and critiques of platform governance. This corresponds to what Liu & McLeod (2019) concluded about the users that often engage in counter-framing by providing alternative perspectives or challenging the original editorial stance. In contrast, replies in Turkey tended either to reiterate the informational content of the posts or to divert the discussion toward moral or hostile commentary unrelated to *Meta*'s policy change.

The findings indicate that user-reply patterns are shaped more by media ecosystems than by national regulatory frameworks or headline tone (SO4). Although the study originally included Turkey, it was excluded from the sentiment analysis due to an insufficient number of statistically significant responses. In the United States and Spain, variations in user sentiment correspond primarily to the ideological orientation and editorial profiles of individual outlets rather than to differences in digital governance models. These results suggest that audience reactions are embedded within media-specific interpretative communities, where prior alignments and loyalties play a more decisive role than national context. In this regard, Spinde et al. (2023) found that the level of hostility expressed in user comments tends to increase in relation to the degree of bias present in the original article. They also provide evidence that the ideological orientation of the media outlet influences the bias embedded in the published content. Consequently, the editorial stance of the outlet emerges as a relevant factor shaping audience reactions, alongside the specific characteristics of the comments themselves (2023).

Software availability

Source code available from: https://github.com/pablopastora/meta-content-moderation-x-pipeline

Archived software available from: https://doi.org/10.5281/zenodo.19355020

License: MIT.

The data collection procedures relied on the following third-party open-source software:

- TweetScraperR (Nieto, A.) – used for scraping media posts from X

https://github.com/agusnieto77/TweetScraperR

- twscrapeR (Nieto, A.) – used for extracting user replies

https://github.com/agusnieto77/twscrapeR

Sentiment analysis was conducted using transformer-based models implemented via the Hugging Face transformers library.

Ethics and consent

Ethical approval and consent were not required.

Data availability

The dataset supporting the conclusions of this article is available at:

https://doi.org/10.5281/zenodo.18431324 (Pastora Estebanez, P., et al, 2026).

Data collection and analysis pipeline for media reactions on X to Meta's content moderation policy. Zenodo.

The dataset contains Tweet IDs and User IDs only. No tweet texts or reply contents are included, in accordance with platform terms of service and ethical restrictions surrounding social media data.

Data are available under the terms of the Creative Commons Attribution 4.0 International license (CC-BY 4.0).

The underlying data to this research cannot be shared due to the ethical and copyright restrictions surrounding social media data. The Methods section contains detailed information to allow replication of the study. Any queries about the methodology should be directed to the corresponding author.

Acknowledgments

I confirm that, in accordance with the F1000 AI Policy, I have used a generative AI tool (ChatGPT, version 5.2, OpenAI) during the preparation of this manuscript. The tool was employed for the translation of certain sections into English and for adapting the text to an appropriate academic tone when required. No content, data, or analyses were generated by the tool; its use was limited solely to linguistic and stylistic support.

References

Adam D: **Does fact-checking work? What the science says.** *Nature.* 2025, January 10.
Publisher Full Text | **Reference Source**

Arun A, Chhatani S, An J, *et al.*: X-posing free speech: Examining the impact of moderation relaxation on online social networks. *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024).*: Association for Computational Linguistics; 2024: 201–211.
Publisher Full Text

Badillo-Matos A, Baldi V, Arteaga F, *et al.*: ***Analysis of the impact of disinformation on political, economic, social and security issues, governance models and good practices: The cases of Spain and Portugal.*** *IBERIFIER.* 2023;
Publisher Full Text

Bakshy E, Messing S, Adamic LA: **Exposure to ideologically diverse news and opinion on Facebook.** *Science.* 2015; **348**(6239): 1130–1132.
PubMed Abstract | **Publisher Full Text**

Bauer M, ECIPESisto E, *et al.*: **What is Wrong with Europe's Shattered Single Market? Lessons from Policy Fragmentation and Misdirected Approaches to EU Competition Policy. In Occasional Papers.** 2023.
Reference Source

Baykürt B: **From circulating liberalism to tech nationalism: THE US. soft power and Silicon Valley.** *Journal of Political Power.* 2022; **15**(3): 456–468.
Publisher Full Text

BBC News Türkçe: *CHP'li Burak Erbay 'Dezenformasyonla Mücadele' yasa teklifini Meclis kürsüsünde çekiçle telefonunu kırarak protesto etti.* BBC; 2022, 5 de octubre).
Reference Source

Bonagiri A, Li L, Oak R, *et al.*: *Towards safer social media platforms: scalable and performant Few-Shot harmful content moderation using large language models.* arXiv (Cornell University); 2025.
Publisher Full Text

Bradshaw S, Howard PN: *The global disinformation order: 2019 global inventory of organised social media manipulation.* Oxford Internet Institute: University of Oxford; 2019.

Candeub A: **Media Ownership Regulation, the First Amendment, and Democracy's Future.** *UC Davis Law Review.* 2008; **41**(41)
Reference Source

Car P: **Fact-checking and content moderation.** *European Parliamentary Research Service.* 2025.
Reference Source

Conte R, Gilbert N, Bonelli G, *et al.*: **Manifesto of computational social science.** *The European Physical Journal Special Topics.* 2012; **214**: 325–346.
Publisher Full Text

Cosentino G, Alikasifoglu B: **Políticas de la posverdad en el Próximo Oriente: los estudios de caso de Siria y Turquía.** *Artnodes.* 2019; **24**: 91–100.
Publisher Full Text

Cuihong C, Liting D: **Evolution of Internet Governance in China: Actors and Paradigms.** *China Quarterly of International Strategic Studies.* 2021; **07**(01): 79–109.
Publisher Full Text

Dickerson C: **Epistemological Threads of American Journalism.** *Pepperdine Policy Review.* 2019; **11**(1).
Reference Source

Edelmann A, Wolff T, Montagne D, *et al.*: **Computational social science and sociology.** *Annual Review of Sociology.* 2020; **46**: 61–81.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Ertuna C, Aşık O: **Journalist YouTubers: How Platformization Transforms Journalism in an Authoritarian System.** *International Journal of Communication.* 2025; **19**: 22–22.
Reference Source

Filgueiras F, Almeida V: **Governance for Digital Technologies.** *Governance for the Digital World.* Cham: Palgrave Macmillan; 2021; pp. 75–104.
Publisher Full Text

Freedom House: *Turkey.* Freedom House; 2024.
Reference Source

Freedom House: *Turkey.* Freedom House; 2025.
Reference Source

Fukuyama F, Grotto A: **Comparative media regulation in the United States and Europe.** *Cambridge University Press eBooks.* 2020; pp. 199–219.
Publisher Full Text

García Madariaga JM, Arasanz Esteban I: **Aprendices explotados: Precarización del trabajo periodístico a través de una lógica abusiva de becas y prácticas.** *Cuadernos de Relaciones Laborales.* 2019; **37**: 49–66.
Publisher Full Text

García-Santamaría JV, Pérez-Serrano MJ: **Grupos de comunicación en España: madurez y profundas transformaciones en un final de ciclo.** *Palabra Clave.* 2020; **23**(4): 1–32.
Publisher Full Text

Gillespie T: *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media.* Yale University Press; 2018.
Publisher Full Text

Gomes AB, Sultan A: **Problematizing content moderation by social media platforms and its impact on digital harm reduction.** *Harm Reduction Journal.* 2024; **21**(194): 2024.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Gorwa R: ***What is platform governance?*** *Information, Communication & Society.* 2019; **22**(6): 854–871.
Publisher Full Text

Gosztonyi G, Gyetván DG, Kovács A: **Theory and practice of social media's content moderation by artificial intelligence in light of European Union's AI Act and Digital Services Act.** *European Journal of Law and Political Science.* 2025; **4**(1): 33–42.
Publisher Full Text

Govindankutty S, Kumar A: **Design and Implementation of Automated Content Moderation Systems in Social Media.** *Integrated Journal for Research in Arts and Humanities.* 2024; **4**(6): 380–402.
Publisher Full Text

Hallin DC, Mancini P, Bretones MT, *et al.*: *Sistemas mediáticos comparados: tres modelos de relación entre los medios de comunicación y la política.* Editorial Hacer; 2008.

Harbi AG: **The role of the US media in the political trajectory of the presidential candidate Donald Trump in 2016: Discourse analysis.** *Jurnal Smart.* 2024; **10**(2): 155–164.
Publisher Full Text

Henton D: **Lessons from Silicon Valley: Governance in a Global City-Region.** *Global City-Regions Trends, Theory, Policy.* 2001; pp. 391–400.
Publisher Full Text

Herman ES, Chomsky N: *Manufacturing consent: the political economy of the mass media.* Pantheon books; 1988.

IG: *FAANG o GAFAM: ¿Son las big tech una buena oportunidad de inversión?.* IG Group; 2021.
Reference Source

Information Technology & Innovation Foundation: *Turkey's content moderation regulation.* ITIF; 2025, September 2.
Reference Source

Jurkowitz M, Gottfried J: *Twitter is the go-to social media site for U.S. journalists, but not for the public.* Pew Research Center; 2022 June 27.
Reference Source

Kalliris K: **Online harm, free speech, and the 'legal but harmful' debate: an interest-based approach.** *Journal of Media Law.* 2024; **16**: 390–416.
Publisher Full Text

Kohl U: **Toxic recommender algorithms: immunities, liabilities and the regulated self-regulation of the Digital Services Act and the Online Safety Act.** *Journal of Media Law.* 2024; **16**: 301–335.
Publisher Full Text

Lazer DMJ, Baum MA, Benkler Y, *et al.*: **The science of fake news.** *Science (New York, N.Y.).* 2018; **359**(6380): 1094–1096.
Publisher Full Text

Liu J, McLeod DM: **Counter-framing effects of user comments. *International.*** *Journal of Communication.* 2019; **13**: 20.
Reference Source

Luna-Reyes LF: **Opportunities and challenges for digital governance in a world of digital participation.** *Information Polity: The International Journal of Government & Democracy in the Information Age.* 2017; **22**(2/3): 197–205.
Publisher Full Text

Majó-Vázquez S, González-Bailón S: **Polarización en las audiencias de los medios en España.** *EsadeEcPol Insight #41.* EsadeEcPol - Center for Economic Policy; 2022.
Reference Source

Nehe Y, Sarak R, Shete R, *et al.*: **Hybrid Moderation Framework for Social Media: Combining AI and Human Expertise.** *International Journal of Scientific Research in Engineering and Management.* 2024; **08**(12): 1–6.
Publisher Full Text

Newman N, Fletcher R, Robertson CT, *et al.*: *Reuters Institute Digital News Report 2024.* Reuters Institute for the Study of Journalism; 2024.
Publisher Full Text

Nieto A: *TweetScraperR: An R package for collecting Twitter/X data via scraping [Software].* GitHub; 2025a.
Reference Source

Nieto A: *twscrapeR: An R package used for extracting replies from X/Twitter [Software].* GitHub; 2025b.
Reference Source

O'Hara K, Hall W: *Four Internets: The Geopolitics of Digital Governance.* Centre for International Governance Innovation; 2018; vol. **206**.
Reference Source

Pariser E: *The filter bubble: What the internet is hiding from you.* Penguin Books; 2012.
Publisher Full Text

Pastora Estebanez P: **Data collection and analysis pipeline for media reactions on X to Meta's content moderation policy (v2.0.0).** *Zenodo.* 2026.
Publisher Full Text

Phillips W, Milner RM: *You are here: A field guide for navigating polarized speech, conspiracy theories, and our polluted media landscape.*: MIT Press; 2021.
Publisher Full Text

Pickard V: *America's battle for media democracy: The triumph of corporate libertarianism and the future of media reform.* Cambridge University Press; 2015.

Reporters without borders: *Turkey. Reporter without bourders* 2024.
Reference Source

Reuters: *U.S. threatens countermeasures on European service providers after EU fines.* Reuters; 2025, December 17.
Reference Source

Rogan J(Host): *#2255—Mark Zuckerberg [Audio podcast episode].* Spotify: In The Joe Rogan Experience; 2025, January 10.
Reference Source

Rone J: **The return of the state? Power and legitimacy challenges to the EU's regulation of online disinformation.** *Power and Authority in Internet Governance.* Routledge; 2021; pp. 171–194.
Publisher Full Text

Sánchez Muñoz O: *La regulación de las campañas electorales en la era digital. Desinformación y microsegmentación en las redes sociales con fines*

electorales. Madrid, Valladolid: Centro de Estudios Políticos y Constitucionales; 2020.

Saka E: ***Social media in Turkey as a space for political battles: AKTrolls and other politically motivated trolling.*** *Middle East Critique.* 2018; **27**(2): 161–177.
Publisher Full Text

Seijas Costa R, Barredo Ibáñez D, Cea Esteruelas N: **Evolutionary regulatory dynamics in a pluralist and polarized journalism landscape: a case study of the normative framework in Spanish media.** *Frontiers in Communication.* 2024; **9**.
Publisher Full Text

Serrano-Puche J: **Credibilidad y confianza en los medios de comunicación: Un panorama de la situación en España.** *Discursos comunicativos persuasivos hoy.* Tecnos; 2017; pp. 427–438.

Spinde T, Richter E, Wessel M, *et al.*: ***What do Twitter comments tell about news article bias? Assessing the impact of news article bias on its perception on Twitter*.** *Online Social Networks and Media.* 2023; **37-38**: 100264.
Publisher Full Text

Thompson A: **Informe Digital Global, Abril 2024. We Are Social Spain.** 2024. abril 29.
Reference Source

Worldometer: *GDP by country.* Worldometer; 2023.
Reference Source

Yücel A: **Hybrid digital authoritarianism in Turkey: The 'Censorship Law' and AI-generated disinformation strategy.** *Turkish Studies.* 2025; **26**(1): 1–27.
Publisher Full Text

Zilinsky J, Pradel F, Theocharis Y, *et al.*: **Content Warning: Public attitudes on content moderation and freedom of expression.** *mediaTUM – the Media and Publications Repository of the Technical University Munich (Technical University Munich).* 2025.
Publisher Full Text

Open Peer Review

Current Peer Review Status: ?

Version 1

Reviewer Report 05 October 2026

https://doi.org/10.21956/openreseurope.24972.r81114

© **2026 AKBAYIR M.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Mustafa AKBAYIR
University of Akdeniz, Antalya, Turkey

Is the work clearly and accurately presented and does it engage with the current literature?

Partly:

The manuscript is generally well structured and written in clear academic English, with objectives, methods and results organised coherently. It engages appropriately with core literature on digital governance models , media systems and content moderation. However, accuracy is undermined by occasional factual errors (most notably the incorrect claim about Meta's user numbers) and minor inconsistencies in methodological description, which prevent a fully affirmative assessment.

Is the study design appropriate and is the work technically sound?

Partly:

The comparative mixed-methods design is appropriate for examining media framing and audience responses across three distinct digital governance contexts, and the combination of engagement metrics, qualitative coding and polarity modelling is technically coherent. Nevertheless, the modest sample of 56 media posts and the insufficient number of Turkish replies for quantitative sentiment analysis reduce statistical power and limit the robustness of cross-national claims, so the work cannot be judged fully technically sound in its present form.

Are sufficient details of methods and analysis provided to allow replication by others?

Yes:

The Methods section describes the main stages of data collection, cleaning, coding and polarity modelling in reasonable detail, and the authors provide open source code and archived software via GitHub and Zenodo. This supports approximate replication.

Are all the source data and materials underlying the results available?

Partly:

The authors have made the analysis code openly available via GitHub and Zenodo, and have deposited a dataset containing Tweet IDs and User IDs. However, the full text of the media posts and user replies is not shared, in line with platform terms of service and ethical restrictions on social-media data. While this is a common and justified limitation, it means that not all source materials underlying the results are fully accessible for independent verification.

If applicable, is the statistical analysis and its interpretation appropriate?

Partly:

The statistical analyses employed (descriptive metrics, chi-square tests of association, and basic correlation/regression checks on polarity scores) are appropriate for the available data and are interpreted with due caution regarding sample limitations. However, the modest number of posts and replies, the exclusion of the Turkish subset from quantitative modelling, and the absence of more robust sensitivity or multiple-comparison checks mean that the statistical component, while generally sound, remains only partially satisfactory.

Are the conclusions drawn adequately supported by the results?

Partly:

The main conclusions regarding cross-national differences in media tone and framing, the predominantly negative user polarity, and the stronger association of sentiment with individual media outlets rather than national context are consistent with the reported results. However, the modest sample size, the exclusion of Turkish replies from quantitative polarity analysis, and the only partial support for some of the original hypotheses mean that the broader claims about media framing as a key mediating mechanism across governance contexts are only partially supported and require more cautious qualification.

Overall assessment and recommendations:

The article offers a timely comparative examination of media framing and user responses on X to Meta's shift from IFCN fact-checking to a community-notes model across three distinct digital-governance contexts. The mixed-methods design is appropriate, the engagement with core literature is adequate, and the principal empirical patterns are coherent with the data presented. Nevertheless, scientific soundness is currently limited by a modest sample, incomplete quantitative coverage of the Turkish case, and occasional factual inaccuracies.
To render the manuscript fully sound the authors must:

- **correct the erroneous statement on Meta user numbers (3.2 million/3.2 billion);**
- **add an explicit Limitations section that quantifies the impact of the small post corpus and the exclusion of Turkish replies from polarity modelling;**
- **map results more systematically onto each hypothesis and qualify the broader claim that media framing operates as the key mediating mechanism across governance regimes.**

Once these essential revisions are completed, the contribution will rest on a firmer empirical and methodological foundation.

Is the work clearly and accurately presented and does it engage with the current literature?
Partly

Is the study design appropriate and is the work technically sound?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

Are all the source data and materials underlying the results available?
Partly

If applicable, is the statistical analysis and its interpretation appropriate?
Partly

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: My research interests focus on media text analysis, critical discourse approaches, and the development of models for news production processes in new media. I particularly examine the ideological framing of news content produced on social media platforms, the construction of sectarian and ethnic discourses, and comparative analyses of traditional media and influencer journalism models. In this context, my study titled “Palestine in the Turkish political Islamist-oriented press: interpreting the 2024 Israel-Iran crisis through social media” (Akbayır & Akbayır, 2026, https://doi.org/10.15178/va.2026.159.e1652), which analyses the social media posts of the Turkish political Islamist-oriented press through quantitative content analysis and critical discourse analysis, serves as a representative example of this line of work.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.