# ANNALES

Anali za istrske in mediteranske študije
Annali di Studi istriani e mediterranei
Annals for Istrian and Mediterranean Studies
Series Historia et Sociologia, 35, 2025, 4

UDK 009 Annales, Ser. hist. sociol., 35, 2025, 4, pp. 393-560, Koper 2025 ISSN1408-5348

UDK 009

ISSN 1408-5348
e-ISSN 2591-1775

# ANNALES

## Anali za istrske in mediteranske študije
## Annali di Studi istriani e mediterranei
## Annals for Istrian and Mediterranean Studies

Series Historia et Sociologia, 35, 2025, 4

KOPER 2025

ISSN 1408-5348
e-ISSN 2591-1775

UDK 009

Letnik 35, leto 2025, številka 4

**Glavni urednik/*Redattore capo/ Editor in chief:*** Darko Darovec

**Odgovorni urednik/*Redattore responsabile/Responsible Editor:*** Salvator Žitko

**Uredniki/*Redattori/Editors:*** Urška Lampe, Boštjan Udovič, Žiga Oman, Veronika Kos

***Gostujoča urednica/Guest Editor/ Editrice ospite:*** Veronika Bajt

**Prevajalka/*Traduttrice*/*Translator:*** Cecilia Furioso Cenci (it.)

**Oblikovalec/*Progetto grafico/ Graphic design:*** Dušan Podgornik, Darko Darovec

**Tisk/*Stampa/Print:*** Založništvo PADRE d.o.o.

**Založnika/*Editori*/*Published by:*** Zgodovinsko društvo za južno Primorsko - Koper / *Società storica del Litorale - Capodistria*© / Inštitut IRRIS za raziskave, razvoj in strategije družbe, kulture in okolja / *Institute IRRIS for Research, Development and Strategies of Society, Culture and Environment / Istituto IRRIS di ricerca, sviluppo e strategie della società, cultura e ambiente*©

**Sedež uredništva/*Sede della redazione/ Address of Editorial Board:*** SI-6000 Koper/*Capodistria,* Garibaldijeva/*Via Garibaldi 18*
***e-mail:*** annaleszdjp@gmail.com, ***internet:*** https://zdjp.si

Redakcija te številke je bila zaključena 30. 12. 2025.

**Sofinancirajo/*Supporto finanziario/ Financially supported by:*** Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost Republike Slovenije (ARIS)

*Annales - Series Historia et Sociologia* izhaja štirikrat letno.

Maloprodajna cena tega zvezka je 11 EUR.

**Naklada/*Tiratura/Circulation:*** 300 izvodov/copie/copies

Revija Annales, Series Historia et Sociologia je vključena v naslednje podatkovne baze / *La rivista Annales, Series Historia et Sociologia è inserita nei seguenti data base / Articles appearing in this journal are abstracted and indexed in*: Clarivate Analytics (USA): Arts and Humanities Citation Index (A&HCI) in/and Current Contents / Arts & Humanities; IBZ, Internationale Bibliographie der Zeitschriftenliteratur (GER); Sociological Abstracts (USA); Referativnyi Zhurnal Viniti (RUS); European Reference Index for the Humanities and Social Sciences (ERIH PLUS); Elsevier B. V.: SCOPUS (NL); Directory of Open Access Journals (DOAJ).

Navodila avtorjem in vsi članki v barvni verziji so prosto dostopni na spletni strani: https://zdjp.si.
*Le norme redazionali e tutti gli articoli nella versione a colori sono disponibili gratuitamente sul sito: https://zdjp.si/it/.*
*The submission guidelines and all articles are freely available in color via website https://zdjp.si/en/.*

Anali za istrske in mediteranske študije - Annali di Studi istriani e mediterranei - Annals for Istrian and Mediterranean Studies

UDK 009 Volume 35, Koper 2025, issue 4 ISSN 1408-5348
e-ISSN 2591-1775

# VSEBINA / *INDICE GENERALE / CONTENTS*

OCENE/*RECENSIONI/REVIEWS*

received: 2025-03-09 DOI 10.19233/ASHS.2025.31

# DIGITAL DISCOURSE DILEMMAS: MODERATING SLOVENIAN DIGITAL LANDSCAPES

*Zoran FIJAVŽ*
Peace Institute, Metelkova ulica 6, 1000 Ljubljana, Slovenia
Jožef Stefan International Postgraduate School, Jamova cesta 39, 1000 Ljubljana, Slovenia
e-mail: zoran.fijavz@mirovni-institut.si

## *ABSTRACT*

*This paper examines content moderation practices in Slovenian digital media organizations, based on interviews with representatives from RTV Slovenija, Mladina, Metropolitan, and Starševski čvek, as well as comment guidelines of the 14 most visited online media and a large forum. Comment standards address socially unacceptable speech beyond illegal content. Larger digital media organizations have developed elaborate in-house systems, while the rest rely on social media platform tools. Sustaining moderation requires significant resources and exacts an emotional toll on moderators. Effective approaches enable flexible moderator responses, utilizing technologies with a range of complexity.*

**Keywords:** content moderation, Digital Services Act, hate speech, media regulation, social media platforms, digital media organizations

## I DILEMMI DEL DISCORSO DIGITALE: MODERARE I PAESAGGI DIGITALI SLOVENI

### *SINTESI*

*Questo articolo esamina le pratiche di moderazione dei contenuti nelle organizzazioni slovene dei media digitali, sulla base di interviste con i rappresentanti di RTV Slovenija, Mladina, Metropolitan e Starševski čvek, nonché delle linee guida per i commenti dei 14 media online più visitati e di un ampio forum. Gli standard per i commenti affrontano discorsi socialmente inaccettabili, oltre i contenuti illegali. Le organizzazioni di media digitali più grandi hanno sviluppato sistemi interni elaborati, mentre le altre si affidano agli strumenti delle piattaforme di social media. Mantenere la moderazione richiede un significativo impegno di risorse e un contributo emotivo ai moderatori. Approcci efficaci consentono risposte flessibili da parte dei moderatori, utilizzando tecnologie con un livello di complessità variabile.*

**Parole chiave:** moderazione dei contenuti, Legge sui servizi digitali, discorso d'odio, regolamentazione dei media, piattaforme di social media, organizzazioni di media digitali

## INTRODUCTION[1]

Moderating digital spaces has gained renewed attention due to the EU's *Digital Services Act* (DSA), requiring large platforms to establish standards for addressing systemic risks like hate speech. While its full effects are yet to be determined, the DSA interacts with existing national norms, such as Slovenia's high hate speech prosecution threshold (Kogovšek Šalamon & Hrvatič, 2024; Vezjak, 2017) and concerns about "awful but lawful" hate speech unaddressed by legal frameworks (Haupt, 2024; Mattheis & Kingdon, 2023). Current large online platform regulation discussions mirror prior debates on moderating website comments and online forums. Before social media regulation debates, news media organizations were recognized as agents against online hate speech (cf. Spletno oko, 2010) and experimented with content moderation approaches (Motl, 2009; Vobič & Poler Kovačič, 2014). The functioning of content moderation in the new regulatory environment is less clear. This study follows up on prior research by examining how selected Slovenian news organizations manage the technical, legal, and labor challenges of moderating online discussions across digital platforms, including social media channels, where legal obligations to respond to user-generated content are often unclear (Korpisaari, 2022).

The next three sections contextualize the study by summarizing the concurrent rise of social media platforms and decline of traditional journalism, the general motivation of digital media organizations to mitigate hate speech, and the existing research and regulatory framework for hate speech in Slovenia. The results section characterizes Slovenian content moderation: its scope, mechanics across platforms, and demands on moderators and their organizations. Finally, the conclusion summarizes the key contributions and reflects on the broader implications for media organizations and digital governance.

### Disruptive innovations, eruptive conversations

Advertising has historically been a key revenue source for news production (Picard, 2011). The news industry was significantly disrupted in the late 2000s by a financial crisis and a transformation of distribution strategies due to the rise of social media and digital advertising (Barrera, 2018; Díaz-Noci, 2020). The former resulted in drastic declines in advertising revenue and readership, mass layoffs of media staff, as well as diminished public trust in traditional media. Print journalism was particularly affected (Barrera, 2018), with US newspaper advertising revenue dropping 47% from 2005 to 2009 (Athey et al., 2013). This was compounded by outlets' challenges with audience monetization in a technologically transformed landscape (Napoli, 2011; Díaz-Noci, 2020). Given direct audience access through platform-based channels, advertisers may choose to bypass journalistic content entirely (Harper, 2017; Sridhar & Sriram, 2015). Subscription-based paywalls proved largely ineffective (Myllylahti, 2014) and the sector commonly responded with outsourcing, cost reductions tied to the precarization of journalists, or pursuing affluent audiences (De Mateo et al., 2010).

Traditional media and digital platforms are thus in a highly asymmetrical relationship in favor of the latter, characterized by competition, conflicting interests, and structural dependency (Kaluža & Slaček Brlek, 2021) as well as contradictory professional values of journalism and technology companies (Russell, 2019). Social media platforms offer limited web traffic and advertising revenue to traditional media (Ju et al., 2014), yet remain a practical necessity (Myllylahti, 2018; Kleis Nielsen & Ganter, 2018). Social media's rise coincided with a drop in labor-intensive content (Shen, 2019), and outlets shifting from hard news to business news and infotainment (Chakravartty & Schiller, 2010), a pattern historically seen during periods of declining advertising and subscription revenues (Angelucci & Cage, 2017). Platform newsfeed algorithms constrained news media's editorial control (Wallace, 2018) with journalists forming "algorithmic folk theories" about content preferred by opaque newsfeed systems (Peterson-Salahuddin & Diakopoulos, 2020) and popular social media content often spills into traditional media (Cage et al., 2020).

Advertising is increasingly commanded by a few large digital platforms with Alphabet and Meta Platforms alone capturing over 60% of global ad revenue (Fuchs, 2018; Graham, 2017).[2] Their core operation has been described as surveillance capitalism: the molding of user data collected beyond service requirements into predictive products (Zuboff, 2019). Google's monetization of user queries set it apart from the many

[1] This research was supported by the Slovenian Research Agency (ARIS) through the research programme P5-0413: *Equality and Human Rights in the Times of Global Governance*, as well as research projects J5-3102: *Hate Speech in Contemporary Conceptualizations of Nationalism, Racism, Gender and Migration* and V5-2330: *Analysis of Online Hate Speech and Disinformation in Slovenia and Development of a Proposal for Action*.

[2] Alphabet, Inc. and Meta Platforms, Inc. are the parent companies of Google and Facebook, respectively.

companies wiped out by the dot-com bubble, following a decade of speculative tech investment in the 1990s (Jayasurya, 2009). Google and Facebook can further be characterized as instances of platform capitalism, marked by strong network effects (exponential utility scaling with more users), the ability to serve as intermediaries (e.g., between brands and audiences on social media), and cross-subsidization of non-profitable services which nevertheless aid data extraction (Srnicek, 2016).

Content moderation on social media thus parallels the tug-of-war between regulators and extractive platforms on topics like online privacy (Srinivasan, 2019), featuring predictable agenda-setting approaches such as lobbying, campaign contributions, regulatory capture, and self-regulation (O'Callaghan & Vivoda, 2013). Concerns over harmful content, such as disinformation, hate speech, and harassment, catalyzed a "techlash" against large platforms and helped consolidate a consensus for increased regulation (Flew et al., 2019). Europe responded by mandating digital platform self-regulation, first in 2017 with the German NetzDG, legitimized by prior waves of anti-refugee hate speech, and more recently with the current EU DSA. While extending online hate speech regulation, these laws heavily rely on user reporting, often by victims lacking support, and leave platform technologies and governance structures intact (Griffin, 2021). NetzDG reportedly reduced overall Twitter/X post toxicity and slightly decreased hate crimes, without altering public opinion on refugees as common hate speech targets (Jiménez Durán et al., 2024). NetzDG's implementation coincided with a shift from report-based to proactive platform removal, accounting for over 90% of Meta's removed hate speech by 2024 (Meta Platforms, 2024b). Digital media organizations, especially comparatively smaller ones in a global view, thus navigate hate speech governance shaped by large platforms and regulators. These are subject to change, as demonstrated by Meta's early 2025 move to defund its fact-checker network and loosen moderation standards (Meta Platforms, 2025).

## Media-managed toxicity

Digital media's interactivity redefined the journalist-audience relationship; media reacted by engaging in comment threads, (rarely) ignoring comments entirely, or implementing moderation systems (Chen & Pain, 2017). Most users hesitate to engage in comment threads with many uncivil comments, especially if such comments start the thread (Lu et al., 2023). The scope of moderation is best captured by broad frameworks such as online toxicity (Wulczyn et al., 2017) and socially unacceptable discourse (Vehovar et al., 2020), encompassing user behaviors from incivility and insults to slurs and threats. Nonetheless, hate speech, generally understood as harmful language directed towards minorities (Papcunová et al., 2023), remains a key concern due to its effects of increased bias towards minority groups (Soral et al., 2018), links to political radicalization (Bilewicz & Soral, 2020), psychological harm to its victims (Wypych & Bilewicz, 2024), and the facilitation of future hate posts (Goel et al., 2023). The scope of hate speech remains contested: "legal" accounts, common in Slovenian practice, limit it to direct incitement to violence (Kogovšek Šalamon & Hrvatič, 2024; Vezjak, 2017), while broader "sociological" interpretations focus on the communicative effects of such speech (Bajt, 2017). The latter further situate hate speech within idiosyncratic socio-cultural norms (Baider, 2020) and performative contexts (Udupa & Pohjonen, 2019), rejecting a pro forma delineation of "regular" and hate speech detached from their wider social settings. Hate speech may promulgate precisely through discursive norm transgression through "fringe" content (Mattheis & Kingdon, 2023). Hate speech is often equated with hostile communication and juxtaposed with civil and polite speech (cf. Ksiazek et al., 2015). While this distinction may be useful for large comment corpora where hostile and discriminatory content overlap, it becomes less useful for complex communicative acts, such as disinformation, that imitate "neutral" news, but can nevertheless promulgate hate speech as an audience response (Hameleers et al., 2022).

The range of these communicative acts is generally covered by norms on content moderation, defined as the monitoring and managing of user-generated content on digital platforms to ensure community standard compliance, prevent harm, and maintain a constructive engagement environment (Gillespie, 2018). It is ubiquitous in online spaces and used by media organizations to preserve their reputation, maintain editorial control, and foster audience engagement (Chen & Pain, 2017). Content moderation operates as a set of interlinked practices and norms in online spaces that extend beyond content removal. For instance, user verification procedures grant distinct levels of user anonymity, which has been linked with a higher prevalence of socially unacceptable speech, potentially through depersonalization and reduced accountability (Bargh, 2002; Cho & Kwon, 2015). Hate speech in a broad sense is often "awful but lawful" (Haupt, 2024; Mattheis & Kingdon, 2023) and can be approached with softer methods like comment section norm-making or rewarding quality engagement (Antoci et al., 2016; Friess et al., 2021; Heinbach et al., 2022; Wolfgang et al., 2020). Digital interfaces directly shape the possible responses of moderators (Jhaver et al., 2023), yet tools provided by major commercial platforms remain limited (Kuo et al., 2023). Facebook, for example, allows automatic comment hiding based on pre-defined comment or author characteristics, such as the presence of images, links, or pre-defined keywords (Meta Platforms, n.d.). However, fine-grained control is currently limited to group administrators and not page administrators, who typically manage the social media presence of media organizations. Group administrators can issue temporary bans, close comment threads, and receive "conflict alerts" which flag escalating exchanges (Meta Platforms, 2024a). Automated methods for harmful content identification have been developed to address large volumes of user-generated content (Piot et al., 2024), but they achieve mixed success beyond formulaic, repetitive texts, require expansive high-quality training datasets, and are often context-specific due to the absence of widely adopted standards or ontologies. Details on models used by social media platforms remain scarce (Gorwa et al., 2020), even under the transparency requirements of the DSA (Meta Platforms, 2024c). Moderators generally experience adverse psychological outcomes which further differ based on their employment status and the specific content moderated (Steiger et al., 2021; Spence et al., 2023).

## Hate speech and media regulation in Slovenia

Slovenian media trends follow global ones but show slower digital marketing uptake, reliance on "catch-all" ad strategies, and sustained focus on TV advertising (Slaček Brlek & Kaluža, 2022). Newspapers responded to falling revenues by cutting labor costs and flexibilizing journalist roles (Bembič & Vobič, 2021; Slaček Brlek & Kaluža, 2022; Vobič, 2013), integrating audience metrics into editorial decisions, increasing advertiser collaboration, or branching into sectors such as event management (Slaček Brlek & Kaluža, 2022). The limited technological investment focused on labor intensification and service diversification (Slaček Brlek & Kaluža, 2022; Slaček Brlek & Tomanić Trivundža, 2019). Facebook is the leading social media platform in Slovenia, used by approximately 72% of individuals aged 15 or above (European Parliament, 2023). Social media platforms are an important distribution channel for Slovenian digital media organizations, which exert negligible influence on platform policies (Kaluža & Slaček Brlek, 2021).

Article 297 of the Slovenian Criminal Law criminalizes hate speech, prohibiting public incitement to hatred, violence or intolerance based on nationality, race, religion, gender or other personal characteristics (KZ-1, 2008). Charges are rare and limited to cases meeting the criterion of "likely disturbance of public order" in spite of available broader legal interpretations (Čufar, 2021; Kogovšek Šalamon & Hrvatič, 2024). Article 8 of the Slovenian Mass Media Act (ZMed, 2006)[3] holds chief media editors liable for the dissemination of content inciting inequality, violence, or hatred. News media are required to establish moderation rules, prohibit hate speech, and respond to it promptly under Articles 16 and 21 of the Code of Slovenian Journalists (Društvo novinarjev Slovenije, 2010). The 2010 Code for Regulating Hate Speech in Slovenian Web Portals, a self-regulatory document accepted by eight major Slovenian online digital media organizations, obliges signatories to implement user verification, content moderation, community reporting, and a warning label on the legal consequences of posting hate speech (Spletno oko, 2010). The Code's initiator, Spletno oko, operated a hate speech reporting center, which was discontinued in 2022 (Vehovar, 2022). The DSA has been in effect since April 2024; yet by September 2024, Slovenian authorities had not issued any takedown orders for illegal content on Facebook (Meta Platforms, 2024c).

Two prior studies examined content moderation practices in Slovenia. Motl (2009) focused on editorial responses to hate speech in Slovenian news outlets, showing most faced daily occurrences and responded by deleting comments, blocking users, disabling comments on sensitive topics,[4] and notifying sanctioned users. Resource constraints were the main reason for avoiding extensive practices, such as default pre-moderation.[5] Moderation presented a double bind of potential legal consequences for insufficient action and negative user responses to moderation (cf. Motl, 2009, 50). The second major study (Vobič & Poler Kovačič, 2014) reviewed media compliance with the Code for Regulating Hate Speech, finding a range of moderation strategies,

[3] A new media law (ZMed-1, 2025) was adopted in 2025.

[4] Primarily črna kronika (Sl.), the section of newspapers or media that reports on crime, accidents, and other tragic or sensational events.

[5] Pre-moderation is the practice of requiring moderators to confirm comments before they are posted. The practice of deleting comments after they are posted is called post-moderation.

including locking comment threads, applying keyword filters, removing or editing comments, periodically erasing entire comment threads, selective pre-moderation, user outreach, supervising repeat offenders, and blocking users. Later studies analyzed hate speech on social media by manual comment annotation, finding about half of Facebook comments on refugee or LGBTIQ+ news contained socially unacceptable speech (Vehovar et al., 2020). Recurring surveys also show the Slovenian public overwhelmingly opposes hate speech (Vehovar, 2023).

## METHODOLOGY

We analyzed documents and conducted interviews with representatives from digital media organizations to describe their content moderation. The document analysis included community guidelines, commenting rules, and other relevant documents from the 14 most visited Slovenian news sites (Semrush, 2024) and a large online forum. Using web traffic ranking, we contacted digital media organizations and conducted four interviews with representatives recognized for their expertise in content moderation by their organizations. Interviews followed a topic list of organizational practices, the role of hate speech in removed content, and implementation challenges. Each interview focused on one setting (website, social media, forum), with questions about others. We also included two news articles (Šuštaršič, 2023; Mekina, 2024) on content moderation at Siol.net and 24ur.com. Interviews were conducted with representatives from:

- **RTV Slovenija**, a high-traffic national public broadcaster, with a focus on their extensive website-based comment sections;
- **Metropolitan**, a multi-brand news website with medium traffic, oriented towards lifestyle and celebrity news, with a focus on content moderation on their social media channels;
- **Mladina**, a weekly news magazine with comparably lower traffic, which disabled website-based comments;
- **Starševski čvek**, a large forum in the Over.net network, owned by Styria Media's Slovenian division. Some forums in the network, such as MedOver.net, are specialized, but Starševski čvek receives a broad range of discussions, including socio-political issues.

A major challenge was the low response rate. Only four of 14 contacted organizations agreed to an interview. Reasons for non-participation are unclear, though one declining representative cited compliance with existing regulation, resource constraints, and the view that social media moderation is the platforms' responsibility. The participating entities still allow limited comparison across key dimensions: organizational scale (large national to specialized media), primary platform focus (website, social media, forum-based moderation), and different operational resources. The results preclude sector-wide generalization, particularly for larger private-sector media like 24ur.com. Based on interviewee reports of past moderator harassment, the author opted to refer to interviewees by organizational affiliation as a precaution.

A further limitation stems from using Semrush's domain-aggregated traffic data for selecting media organizations. The locally established metric repository MOSS provides clearer subdomain traffic metrics for Slovenian websites. Either source would include major websites, but subdomains may represent entirely separate media brands, as exemplified by Svet24.si with subdomains for Radio Celje, Radio Ptuj, Dolenjski list, Primorske novice and Vestnik, among others. Conversely, MOSS is opt-in, excluding some organizations listed on Semrush, such as Nova24TV. However, even with detailed subdomain traffic data, affiliated brands may share staff and resources, as in the case of content moderation at Metropolitan.

## RESULTS

Three key themes emerge from our study. Firstly, while hate speech regulation is a prominent policy concern, digital media organizations' moderation extends beyond strictly illegal content. Secondly, the technological affordances available to moderators vary significantly between in-house systems and social media platforms, influencing both the scope and effectiveness of content moderation efforts. Thirdly, content moderation imposes substantial resource demands on digital media organizations, as large, irregular, and often affectively charged comment volumes create logistical and psychological challenges for moderators and organizations.

### Language that is offensive or off-limits?

Interview and document data show content moderation aims to mitigate legal and reputational risks while enforcing broader norms like comment civility. Nearly all examined digital media organizations prohibit hate speech in their comment guidelines. 24ur.com explicitly prohibits hate speech, moderating comments deemed "racist, sexist, homophobic" (Pro Plus, 2022). MMC RTV Slovenija bans hate speech, incitement to violence, and intolerance, warning users of criminal responsibility (MMC RTV Slovenija, 2014). N1 rejects comments with hate speech based on national, racial, religious, or other intolerance (N1, 2018). Žurnal24, Slovenske novice, and Delo also ban discriminatory comments (Žurnal24, n.d.; Delo, 2021; Slovenske novice, 2022). Nova24TV, a far-right outlet, prohibits calls to violence (Nova24TV, 2019) and while its guidelines omit hate speech, it is explicitly prohibited by Disqus, the site's third-party comment service (Disqus, n.d.). Forwarding comments to law enforcement is rare: 24ur.com reported fewer than five incidents in seven years (Mekina, 2024) and RTV Slovenija annually forwards one or two illegal hate speech instances. Views on hate speech differ among the digital media organizations studied. RTV Slovenija and Mladina focus on its harm to minorities, while Starševski čvek distinguishes between negative opinions towards minorities and background-motivated verbal abuse, banning the latter. Metropolitan defines hate speech as incitement to violence but also moderates comments targeting individuals (public figures or other commentators).

Comment guidelines often prohibit hate speech within a broader civility norm encompassing inappropriate (though not necessarily illegal) content like offensive or vulgar material, harassment, provocation, "shouting" (e.g., fully capitalized comments), and disruptive behaviors like spamming or off-topic posts. RTV Slovenija, for instance, moderates relatively innocuous comments like grammatical corrections, technical support requests, or story leads, as dedicated email channels exist for these (MMC RTV Slovenija, 2014). Another banned content category concerns commercial content, such as advertising and copyrighted materials. Additionally, non-Slovenian language comments are moderated to comply with the Public Use of the Slovene Language Act, which mandates Slovenian in public communication, including news comments (ZJRS, 2004). Digital media organizations cite various content moderation motivations. Interviews show that Metropolitan emphasizes maintaining brand identity. General negative comments are more acceptable in celebrity news, but inappropriate for brands focusing on health and spirituality. The RTV interviewee explained that moderation aims to foster high-quality public discourse and deliberative practices in line with its public service mission. Similarly, Mladina and Siol.net cited adhering to professional journalistic standards for closing unmoderated comment sections (Šuštaršič, 2023).

In contrast to policy debates that weigh free speech against hate speech, we find content moderation enforces a wide set of norms connecting to civility, relevance, commercial interest, and linguistic identity, which together shape the practical regulation of "awful but lawful" content well beyond simple removal.

## Content moderation mechanics

The materials analyzed show that online spaces differ in regulatory standards and technological affordances. Content moderation practices are bifurcated by setting. While major digital media organizations use extensive in-house moderation, less-visited ones often delegate comments to social media and some support website comments with the Facebook Comment plug-in, which receives a negligible number of comments. Comment guidelines and regulations primarily focus on website comments, rarely detailing their applicability to social media channels. Shifting comments to social media is an attractive solution for smaller digital media organizations, as it reduces legal liability and in-house costs while engaging existing users on social media platforms. Yet, compared to in-house systems, social media moderation is significantly constrained, lacking simple moderation features such as the ability to disable comments on individual posts (e.g. on Facebook), issue temporary bans, and archive removed comments for legal and appeal purposes. Table 1 summarizes key content moderation characteristics of Slovenian digital media, using data from interviews and document analysis.

RTV Slovenija employs a complex moderation framework. Moderators work two shifts, alternating between pre-moderation, post-moderation, and disabling comments on individual news items as required. Post-moderation is the default and topics like crime news are categorically closed to comments. Commenting is disabled overnight due to the discontinuation of night shifts. Pre-moderation applies to contested topics (e.g., armed conflicts, LGBTQ+ issues) with high potential for editorial breaches. Exceptional cases (e.g., the COVID-19 pandemic) warrant extra in-house staff for moderation. Commenting rules and a community reporting option are linked above the comment box. Users breaching comment standards face gradual sanctions: comment removal, warnings, temporary bans, or permanent account blocks. Moderation decisions may be appealed to the RTV Slovenija Ombudsman. A specialized machine learning model complements human review. A social media editor enforces comparable

***Table 1: Overview of content moderation practices at examined Slovenian digital media organizations.***

| *Media organization* | *Web comment support* | *Key moderation approaches* | *Commenter verification* | *Appeal to moderation decisions* |
|---|---|---|---|---|
| rtvslo.si | Own system | Default post-moderation<br>Topical pre-moderation<br>Comments off overnight<br>AI flagging<br>Comment deletions<br>Community flagging<br>User warnings<br>Temporary, permanent ban | Registration | RTV Slovenia ombudsman |
| 24ur.com | Own system | Post-moderation<br>AI flagging<br>Comments off overnight<br>Keyword filters<br>Community flagging<br>Temporary, permanent ban | Registration | Appeals email |
| siol.net | Disabled | — | — | — |
| zurnal24.si | Own system | Post-moderation<br>Permanent user ban<br>Community flagging | Registration | Not stated |
| n1info.si | Own system | Full pre-moderation | None | Not stated |
| slovenskenovice.si | Disabled | Post-moderation<br>Comment edits or deletion<br>Community reporting<br>Permanent user bans | — | Not stated |
| svet24.si | Disabled | — | — | — |
| metropolitan.si | FB Comments plugin (website) | For social media:<br>Post-moderation<br>Comment deletion & user ban (simultaneous) | FB account (website) | Not stated |
| delo.si | Own system | Post-moderation<br>Comment edits or deletion, Permanent user bans | Registration and subscription | Not stated |
| nova24tv.si | Disqus | Comment deletion<br>Community flagging<br>User bans | Disqus<br>Registration | Not stated |
| dnevnik.si | Disabled | — | — | — |
| necenzurirano.si | FB Comments plugin | Not specified | FB account | Not stated |
| reporter.si | FB Comments plugin | Post-moderation<br>Comment edits or deletion<br>Temporary & permanent user bans | FB account | Not stated |
| vecer.com | Web comments mentioned in ToS, not visible on website | Post-moderation<br>Comment deletion<br>User bans | Unknown | Not stated |
| Starševski čvek | Forum | Post-moderation<br>Edit, lock, or delete comments<br>Community flagging<br>User bans | Possible, not required | Appeals email |

standards on social platforms and produces additional social media content.

Metropolitan's social media managers handle moderation, with a daily focus on Facebook, which is their largest channel with the highest comment volume. It is conducted in Meta Business Suite with comment deletion and user blocks. The platform does not provide archiving of removed content. Moderation effectiveness is attributed to good internal brand awareness and coordination. Brand editors will notify social media managers of inappropriate content and provide moderation support if needed, but comment surges are comparatively rare. While a dedicated community manager would admittedly be ideal, they found a collaborative approach to be an effective practical solution.

Starševski čvek edits or deletes posts and locks threads with an option to appeal moderation decisions via a provided email (Starševski čvek, 2018). Manual review of active forum topics is the primary approach, and community reports and keyword alerts are available. One full-time and two part-time moderators conduct this work. Deleted comments are archived for potential legal and law enforcement inquiries. Users have high anonymity with changeable usernames, but moderators can identify them with internal tools. Common sanctions include comment deletions and blocks. Direct outreach to users was discontinued as the forum grew. Spam remains a noticeable problem. Context is considered key for moderation decisions, and AI-supported moderation is viewed with both interest and skepticism.

Mladina and Siol.net disabled website comments around 2015 and 2023, respectively (Šuštaršič, 2023). Both lacked resources and staff for effective moderation of comments, which rarely contained valuable insights or meaningful discussions. Siol.net attempted moderation by deleting comments and banning users, but with subpar results (Šuštaršič, 2023). Ultimately, both outlets concluded that inadequately moderated comments constitute harm and go against journalistic ethics. After disabling comments, Mladina noted a slight decrease in web traffic while Siol.net observed no significant change (Šuštaršič, 2023).

### Operational and psychological demands of moderation

The RTV interviewee characterized content moderation as a "dirty job," citing challenges from fluctuating comment volumes and the emotional toll of moderation. RTV Slovenija receives three to five thousand comments daily, particularly on emotionally charged topics such as armed conflicts, natural disasters, and contested social issues. A 24ur.com editor estimates the website annually receives two million comments (Mekina, 2024). On Starševski čvek, inactive forum threads often resurface through organic search.

Moderation is often emotionally taxing due to the potentially disturbing content as well as negative reactions from users facing moderation. RTV Slovenija opted to anonymize moderators after threats and reported difficulties with recruiting for the role. Starševski čvek's moderators often face personal attacks and discontent over their decisions, noting the work requires "strong nerves". Before closing web comments, Siol.net's editor faced ongoing censorship accusations and threats of legal action for moderation (Šuštaršič, 2023).

## CONCLUSION

Content moderation extends beyond illegal content removal to a range of socially unacceptable discourse from incivility, defamatory content, and commercial content to hate speech. The work of moderators is ultimately determined by the total set of standards. The locus and frequency of contentious speech depend on the topic, with overt hate speech against social groups concentrated in news that references them, and interpersonal hostility more prevalent in celebrity or lifestyle coverage.

Moderation capabilities are shaped by their operational context and bifurcated in Slovenia. Larger organizations like RTV Slovenija have developed sophisticated in-house systems with nuanced controls, such as comment thread off-switches and custom AI models. Smaller ones tend to shift user comments to social media platforms where the tooling is limited.

This practical aspect of moderation is outside the scope of the DSA as a flagship social media regulatory framework that establishes a range of reporting and risk assessment obligations. The implementation of these obligations is, furthermore, marked by a near-complete lack of action against illegal content in Slovenia according to the first DSA reports. This occurs even as local media entities are subjected to comparatively higher legal scrutiny, creating an incentive for moving user comments to social media platforms.

Flexibility is key in moderation strategies, as comment volumes can fluctuate heavily. Flexible responses can be achieved by modulating the degree of user interaction with comment sections or managing moderation demands within an organization when required. Content moderation can be psychologically taxing due to repeated exposure to disturbing content, negative user reactions and variable comment volumes, which can lead to difficulties with recruiting personnel. Long-term maintenance of moderation processes is a significant organizational challenge and entangled with the broader financial and operational capacities of digital media organizations.

Our sample included organizations of various sizes, but the findings may not generalize sector-wide due to a low interview response rate and the absence of larger private and smaller local media. The self-reported nature of interviews and guideline documents may omit informative aspects of moderation in practice. Future research would benefit from a larger interview sample, using multiple web traffic sources like Semrush or MOSS to identify relevant entities, and from employing mixed methods, such as combining interviews with direct observation of moderation or computational analysis. Finally, while we describe common barriers across moderation settings, future research could examine the setting-specific user expectations, communication norms, regulative requirements, and moderation practices.

# DILEME DIGITALNEGA DISKURZA: MODERIRANJE SLOVENSKIH DIGITALNIH KRAJIN

*Zoran FIJAVŽ*
Mirovni inštitut, Metelkova ulica 6, 1000 Ljubljana, Slovenija
Mednarodna podiplomska šola Jožefa Stefana, Jamova cesta 39, 1000 Ljubljana, Slovenija
e-mail: zoran.fijavz@mirovni-institut.si

*POVZETEK*

*Članek analizira spletno moderiranje v Sloveniji v luči novih regulacij, kot je Akt o digitalnih storitvah, upoštevaje asimetrično soodvisnost med družbenimi platformami in slovenskimi medijskimi organizacijami. Temelji na analizi dokumentov, kot so medijska pravila, pogoji uporabe, relevantnih novic in poročil družbenih platform, in štirih intervjujih s predstavnicami_ki RTV Slovenije, Metropolitana, Mladine in Starševskega čveka. Obiskanost spletnih prostorov močno kroji način moderiranja: večji mediji ali forumi moderirajo z lastno tehnološko infrastrukturo, manjše organizacije pa se z uporabniškimi vsebinami srečujejo pretežno na družbenih omrežjih, povečini na Facebooku, in moderirajo z omejenimi orodji, ki jih ta ponujajo. Moderiranje uporabniških vsebin na družbenih omrežjih ostaja izbirno, hkrati pa še ni znakov efektivne regulacije digitalnih platform preko novega evropskega Akta o digitalnih storitvah. Rešitve večjih organizacij za moderiranje lastnih strani v primerjavi z orodji družbenih platform ponujajo večjo fleksibilnost z možnostmi večstopenjskega sankcioniranja, nadzora na ravni posamezne novice in arhiviranja za primere pritožbenih ali pravnih postopkov. Pravila komentiranja in izkušnje moderatorjev pokažejo, da moteče, vendar nekaznive vsebine predstavljajo znaten del moderiranja, cilj katerega je ne le izpolnitev regulativnih zahtev, temveč tudi ohranjanje konstruktivnega diskurza in zaščita ugleda organizacij. Moderiranje za analizirane organizacije predstavlja znatno porabo omejenih virov. Narava moderiranih vsebin in pogosti negativni odzivi uporabnikov, ki v omejenih primerih preidejo v verbalno nasilje, lahko predstavljajo čustveno obremenitev za moderatorke_je. V kontekstu razprav o reguliranju uporabniških vsebin na družbenih platformah dodatno pokažemo, da so medijske hiše in forumi za namene moderiranja na lastnih spletnih že razvila odgovore na mnoge izzive moderiranja uporabniških vsebin.*

**Ključne besede:** moderiranje uporabniških vsebin, Akt o digitalnih storitvah, sovražni govor, družbena omrežja, regulacija medijev, digitalne medijske organizacije

## SOURCES AND BIBLIOGRAPHY

**Angelucci, Charles & Julia Cagé (2017):** Newspapers in Times of Low Advertising Revenues. American Economic Journal: Microeconomics, 11, 3, 319–364.

**Antoci, Angelo, Delfino, Alexia, Paglieri, Fabio, Panebianco, Fabrizio & Fabio Sabatini (2016):** Civility vs. Incivility in Online Social Interactions: An Evolutionary Approach. PloS One, 11, 11, Article e0209899.

**Athey, Susan, Calvano, Emilio & Joshua Gans (2013):** The Impact of the Internet on Advertising Markets for News Media. Washington, National Bureau of Economic Research.

**Baider, Fabienne (2020):** Pragmatics Lost? Overview, Synthesis and Proposition in Defining Online Hate Speech. Pragmatics and Society, 11, 2, 196–217.

**Bajt, Veronika (2017):** Sovražni govor kot spodbuda kritičnega delovanja. V: Splichal, Slavko (ur.): Zagovor javnosti: Med svobodo izražanja in sovražnim govorom. Ljubljana, Slovenska akademija znanosti in umetnosti, 70–77.

**Bargh, John (2002):** Beyond Simple Truths: The Human-Internet Interaction. Journal of Social Issues, 58, 1, 1–8.

**Barrera, Carlos (2018):** Las Encrucijadas de Los Medios de Comunicación En La Crisis de 2008-2014: ¿declive o Transformación Del Cuarto Poder?. Historia Actual Online, 47, 3, 79–90.

**Bembič, Branko & Igor Vobič (2021):** Koalicije preživetja v obdobju zatona časopisne industrije: Študija primera. Javnost - The Public, 28, S81–S102.

**Bilewicz, Michał & Wiktor Soral (2020):** Hate Speech Epidemic. The Dynamic Effects of Derogatory Language on Intergroup Relations and Political Radicalization. Political Psychology, 41, S1, 3–33.

**Cagé, Julia, Hervé, Nicolas & Béatrice Mazoyer (2020):** Social Media and Newsroom Production Decisions. SSRN Electronic Journal.

**Chakravartty, Paula & Dan Schiller (2010):** Neoliberal Newspeak and Digital Capitalism in Crisis. International Journal of Communication, 4, 670–692.

**Chen, Gina Masullo & Paromita Pain (2017):** Normalizing Online Comments. Journalism Practice, 11, 7, 876–892.

**Cho, Daegon & Kyounghee Hazel Kwon (2015):** The Impacts of Identity Verification and Disclosure of Social Cues on Flaming in Online User Comments. Computers in Human Behavior, 51, 363–372.

**Čufar, Kristina (2021):** Legal Aspects of Content Moderation on Social Networks in Slovenia. In: Wielec, Marcin (ed.): The Impact of Digital Platforms and Social Media on the Freedom of Expression and Pluralism. Budapest, Ferenc Mádl Institute of Comparative Law – Central European Academic Publishing, 175–216.

**Delo (2021):** Splošni pogoji uporabe spletnih strani medijske hiše Delo d.o.o. (published 10.2021). https://web.archive.org/web/20240418190640/https://trgovina.delo.si/pogoji-uporabe-spletnega-mesta/ (last access: 2024-04-18).

**De Mateo, Rosario, Bergés, Laura & Anna Garnatxe (2010):** Crisis, What Crisis? The Media: Business and Journalism in Times of Crisis. tripleC: Communication, Capitalism & Critique, 8, 2, 251–274.

**Díaz-Noci, Javier (2020):** Cómo Los Medios Afrontan La Crisis: Retos, Fracasos y Oportunidades de La Fractura Digital. El Profesional de La Información, 28, 6, Article e280625.

**Disqus (n.d.):** Terms of Service. https://web.archive.org/web/20240418190302/https://help.disqus.com/en/articles/1717102-terms-of-service#publisher-terms-of-service-agreement (last access: 2024-04-18).

**Društvo novinarjev Slovenije (2010):** Kodeks novinarjev Slovenije. https://novinar.com/drustvo-novinarjev-slovenije/o-nas/dokumenti/kodeks/ (last access: 2024-11-14).

**European Parliament (2023):** Flash Eurobarometer FL012EP: Media & News Survey 2023. https://data.europa.eu/data/datasets/s3153_fl012ep_eng?locale=en (last access: 2024-11-03).

**Flew, Terry, Martin, Fiona & Nicolas Suzor (2019):** Internet Regulation as Media Policy: Rethinking the Question of Digital Communication Platform Governance. Journal of Digital Media & Policy, 10, 1, 33–50.

**Friess, Dennis, Ziegele, Marc & Dominique Heinbach (2021):** Collective Civic Moderation for Deliberation? Exploring the Links between Citizens' Organized Engagement in Comment Sections and the Deliberative Quality of Online Discussions. Political Communication, 38, 5, 624–646.

**Fuchs, Christian (2018):** The Online Advertising Tax. London, University of Westminster Press.

**Gillespie, Tarleton (2018):** Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media. New Haven, Yale University Press.

**Goel, Vasu, Sahnan, Dhruv, Dutta, Subhabrata, Bandhakavi, Anil & Tanmoy Chakraborty (2023):** Hatemongers Ride on Echo Chambers to Escalate Hate Speech Diffusion. PNAS Nexus, 2, 3, pgad041.

**Gorwa, Robert, Binns, Reuben & Christian Katzenbach (2020):** Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance. Big Data & Society, 7, 1, 1–15.

**Graham, Rosie (2017):** Google and Advertising: Digital Capitalism in the Context of Post-Fordism, the Reification of Language, and the Rise of Fake News. Palgrave Communications, 3, 1, 1–19.

**Griffin, Rachel (2021):** New School Speech Regulation and Online Hate Speech: A Case Study of Germany's NetzDG. SSRN Electronic Journal.

**Hameleers, Michael, van der Meer, Toni & Rens Vliegenthart (2022):** Civilized Truths, Hateful Lies? Incivility and Hate Speech in False Information – Evidence From Fact-Checked Statements in the US. Information, Communication & Society, 25, 11, 1596–1613.

**Harper, Tauel (2017):** The Big Data Public and Its Problems: Big Data and the Structural Transformation of the Public Sphere. New Media & Society, 19, 9, 1424–1439.

**Haupt, Claudia E. (2024):** Curbing Hate Speech Online: Lessons from the German Network Enforcement Act (NetzDG). SSRN Electronic Journal.

**Heinbach, Dominique, Wilms, Lena & Marc Ziegele (2022):** Effects of Empowerment Moderation in Online Discussions: A Field Experiment with Four News Outlets [Conference presentation]. Paris, 72nd Annual Conference of the International Communication Association.

**Jayasurya, Gautam (2009):** A Critical Analysis of the Google Story. SSRN Electronic Journal.

**Jhaver, Shagun, Zhang, Alice Qian, Chen, Quan Ze, Natarajan, Nikhila, Wang, Ruotong & Amy X. Zhang (2023):** Personalizing Content Moderation on Social Media: User Perspectives on Moderation Choices, Interface Design, and Labor. Proceedings of the ACM on Human-Computer Interaction, 7, CSCW2, 1–33.

**Jiménez Durán, Rafael, Müller, Karsten & Carlo Schwarz (2024):** The Effect of Content Moderation on Online and Offline Hate: Evidence from Germany's NetzDG. SSRN Electronic Journal.

**Ju, Alice, Jeong, Sun Ho & Hsiang Iris Chyi (2014):** Will Social Media Save Newspapers? Examining the Effectiveness of Facebook and Twitter as News Platforms. Journalism Practice, 8, 1, 1–17.

**Kaluža, Jernej & Sašo Slaček Brlek (2021):** Zapleteno je: protipolno razmerje med globalnimi digitalnimi platformami in slovenskimi mediji. Javnost – The Public, 28, S23–S42.

**Kleis Nielsen, Rasmus & Sarah Anne Ganter (2018):** Dealing with Digital Intermediaries: A Case Study of the Relations between Publishers and Platforms. New Media & Society, 20, 4, 1600–1617.

**Kogovšek Šalamon, Neža & Sergeja Hrvatič (2024):** The Impact of the Slovenian Supreme Court 'Hate Speech' Decision on Prosecutorial Practice. In: Hate Speech Intersections with Nationalism, Racism, Gender and Migration: Peace Institute International Symposium, Abstracts. Ljubljana, Peace Institute. https://www.mirovni-institut.si/wp-content/uploads/2024/07/sovrag-simpozij-program-_abstracts.pdf (last access: 2024-09-12).

**Korpisaari, Päivi (2022):** From Delfi to Sanchez – When Can an Online Communication Platform Be Responsible for Third-Party Comments? An Analysis of the Practice of the ECtHR and Some Reflections on the Digital Services Act. Journal of Media Law, 14, 2, 352–377.

**Ksiazek, Thomas B., Peer, Limor & Andrew Zivic (2015):** Discussing the News: Civility and Hostility in User Comments. Digital Journalism, 3, 6, 850–870.

**Kuo, Tina, Hernani, Alicia & Jens Grossklags (2023):** The Unsung Heroes of Facebook Groups Moderation: A Case Study of Moderation Practices and Tools. Proceedings of the ACM on Human-Computer Interaction, 7, CSCW1, 1–38.

**KZ-1 (2008):** Kazenski zakonik. Uradni list RS, št. 55/08. https://pisrs.si/Pis.web/pregledPredpisa?id=ZAKO5050 (last access: 2025-11-28).

**Lu, Shuning, Liang, Hai & Gina M. Masullo (2023):** Selective Avoidance: Understanding How Position and Proportion of Online Incivility Influence News Engagement. Communication Research, 50, 4, 387–409.

**Mattheis, Ashley A. & Ashton Kingdon (2023):** Moderating Manipulation: Demystifying Extremist Tactics for Gaming the (Regulatory) System. Policy & Internet, 15, 4, 478–497.

**Mekina, Borut (2024):** Ukinimo komentarje. Mladina, April 19, 2024. https://www.mladina.si/232156/ukinimo-komentarje (last access: 2024-11-05).

**Meta Platforms (2024a):** Announcing New Tools and Features to Nurture Your Community. https://www.facebook.com/community/whats-new/new-tools-features-nurture-community/ (last access: 2024-11-05).

**Meta Platforms (2024b):** Community Standards Enforcement Report: Hate Speech. https://transparency.meta.com/reports/community-standards-enforcement/hate-speech/facebook/ (last access: 2024-11-05).

**Meta Platforms (2024c):** Regulation (EU) 2022/2065 Digital Services Act Transparency Report for Facebook. https://transparency.meta.com/sr/dsa-transparency-report-apr2024-facebook (last access: 2024-11-05).

**Meta Platforms (2025):** More Speech and Fewer Mistakes. https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/ (last access: 2025-01-10).

**Meta Platforms (n.d.):** Manage Comments with Moderation Assist for Pages and Professional Mode. https://web.archive.org/web/20241105151254/https://www.facebook.com/help/1011133123133742/ (last access: 2024-11-05).

**MMC RTV Slovenija (2014):** Standardi in pravila komuniciranja na spletnem mestu rtvslo.si. https://web.archive.org/web/20240418162601/https://www.rtvslo.si/uporaba/standardi-in-pravila-komuniciranja-na-spletnem-mestu-rtvslo-si/590527 (last access: 2024-04-18).

**Motl, Andrej (2009):** Sovražni govor v slovenskih medijih na spletu. Ljubljana, Fakulteta za družbene vede.

**Myllylahti, Merja (2014):** Newspaper Paywalls—the Hype and the Reality: A Study of How Paid News Content Impacts on Media Corporation Revenues. Digital Journalism, 2, 2, 179–194.

**Myllylahti, Merja (2018):** An Attention Economy Trap? An Empirical Investigation into Four News Companies' Facebook Traffic and Social Media Revenue. Journal of Media Business Studies, 15, 4, 237–253.

**N1 (2018):** Pravila uporabe (last update: May 25, 2018). https://web.archive.org/web/20231127153112/https://n1info.si/pravila-uporabe/ (last access: 2024-04-18).

**Napoli, Philip M. (2011):** Audience Evolution: New Technologies and the Transformation of Media Audiences. New York, Columbia University Press.

**Nova24TV (2019):** Pravila komentiranja na spletnem portalu Nova24TV https://web.archive.org/web/20240418185654/https://nova24tv.si/obvestilo-pravila-komentiranja-na-spletnem-portalu-nova24tv/ (last access: 2024-04-18).

**O'Callaghan, Terry & Vlado Vivoda (2013):** How Global Companies Make National Regulations. In: Mikler, John (ed.): The Handbook of Global Companies. Chichester, Wiley & Sons, 155–172.

**Papcunová, Jana, Martončik, Marcel, Fedáková, Denisa, Kentoš, Michal, Bozogáňová, Miroslava, Srba, Ivan, Moro, Robert, Pikuliak, Matúš, Šimko, Marián & Matúš Adamkovič (2023):** Hate Speech Operationalization: A Preliminary Examination of Hate Speech Indicators and Their Structure. Complex & Intelligent Systems, 9, 3, 2827–2842.

**Peterson-Salahuddin, Chelsea & Nicholas Diakopoulos (2020):** Negotiated Autonomy: The Role of Social Media Algorithms in Editorial Decision Making. Media and Communication, 8, 3, 27–38.

**Picard, Robert G. (2011):** The Economics and Financing of Media Companies: Second Edition. New York, Fordham University Press.

**Piot, Paloma, Martín-Rodilla, Patricia & Javier Parapar (2024):** MetaHate: A Dataset for Unifying Efforts on Hate Speech Detection. Proceedings of the International AAAI Conference on Web and Social Media, 18, 1, 2025–2039.

**Pro Plus (2022):** Pravila za objavo komentarjev na portalih POP TV in Kanala A (last update: November 18, 2022). https://web.archive.org/web/20240418161936/https://images.24ur.com/media/document/Nov2022/62860264.pdf (last access: 2024-04-18).

**Russell, Frank Michael (2019):** The New Gatekeepers: An Institutional-Level View of Silicon Valley and the Disruption of Journalism. Journalism Studies, 20, 5, 631–648.

**Shen, Neil (2019):** The Newspaper Industry in a Changing Landscape: The Shift in News Content of Various Newspapers as a Response to the Rise of Social Media. Network and Communication Technologies, 5, 1, 1–10.

**Semrush (2024):** Top Websites in Slovenia by Traffic: March 2024. https://web.archive.org/web/20240418233032/https://www.semrush.com/trending-websites/si/all (last access: 2024-04-18).

**Slaček Brlek, Sašo & Jernej Kaluža (2022):** Survival Strategies of Digital News Media in the Changing Market Environment. In: Manninen, Ville J. E., Niemi, Marik K. & Anthony Ridge-Newman (eds.): Futures of Journalism: Technology-stimulated Evolution in the Audience-News Media Relationship. Cham, Palgrave Macmillan, 35–48.

**Slaček Brlek, Sašo & Ilija Tomanić Trivundža (2019):** Algoritmizacija nacionalne tiskovne agencije: primer STA. Javnost – The Public, 26, S62–S81.

**Slovenske novice (2022):** Pogoji uporabe (last update: March 1, 2022). https://web.archive.org/web/20240418184613/https://www.slovenskenovice.si/razno/pogoji-uporabe/ (last access: 2024-04-18).

**Soral, Wiktor, Bilewicz, Michał & Mikołaj Winiewski (2018):** Exposure to Hate Speech Increases Prejudice Through Desensitization. Aggressive Behavior, 44, 2, 136–146.

**Spence, Ruth, Bifulco, Antonia, Bradbury, Paula, Martellozzo, Elena & Jeffrey DeMarco (2023):** The Psychological Impacts of Content Moderation on Content Moderators: A Qualitative Study. Cyberpsychology: Journal of Psychosocial Research on Cyberspace, 17, 4, Article 8.

**Spletno oko (2010):** Kodeks regulacije sovražnega govora na spletnih portalih. https://www.spletno-oko.si/sites/default/files/kodeks_oblikovan_0.pdf (last access: 2024-11-14).

**Srinivasan, Dina (2019):** The Antitrust Case Against Facebook: A Monopolist's Journey Towards Pervasive Surveillance in Spite of Consumers' Preference for Privacy. Berkeley Business Law Journal, 16, 1, 39–101.

**Sridhar, Shrihari & Srinivasaraghavan Sriram (2015):** Is Online Newspaper Advertising Cannibalizing Print Advertising? Quantitative Marketing and Economics, 13, 283–318.

**Srnicek, Nick (2016):** Platform Capitalism. Cambridge, Polity Press.

**Starševski čvek (2018):** Informacije in pravila. https://forum.over.net/informacije-in-pravila/ (last access: 2024-04-29).

**Steiger, Miriah, Bharucha, Timir J., Venkatagiri, Sukrit, Riedl, Martin J. & Matthew Lease (2021):** The Psychological Well-Being of Content Moderators: The Emotional Labor of Commercial Moderation and Avenues for Improving Support. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 1–14.

**Šuštaršič, Mihael (2023):** Komentiranje brez smisla. Siol.net, May 24, 2023. https://siol.net/mnenja/kolumne/komentiranje-brez-smisla-607315 (last access: 2024-11-05).

**Udupa, Sahana & Matti Pohjonen (2019):** Extreme Speech and Global Digital Cultures: Introduction. International Journal of Communication, 13, 3049–3067.

**Vehovar, Vasja (2022):** Sovražni govor: pregled anketnih raziskav in prijavnih točk. Ljubljana, Fakulteta za družbene vede.

**Vehovar, Vasja (2023):** Sovražni govor 2023: Analiza anketne raziskave. Ljubljana, Fakulteta za družbene vede.

**Vehovar, Vasja, Povž, Blaž, Fišer, Darja, Ljubešić, Nikola, Šulc, Ajda & Dejan Jontes (2020):** Družbeno nesprejemljivi diskurz na facebookovih straneh novičarskih portalov. Teorija in Praksa, 57, 2, 622–645.

**Vezjak, Boris (2017):** Interpretacije 297. člena Kazenskega zakonika, opredelitev in pregonljivost sovražnega govora. V: Splichal, Slavko (ed.): Zagovor javnosti: Med svobodo izražanja in sovražnim govorom. Ljubljana, Slovenska akademija znanosti in umetnosti, 77–89.

**Vobič, Igor (2013):** Journalism and the Web: Continuities and Transformations at Slovenian Newspapers. Ljubljana, Ljubljana University Press, Faculty of Social Sciences.

**Vobič, Igor & Melita Poler Kovačič (2014):** Keeping Hate Speech at the Gates: Moderating Practices at Three Slovenian News Websites. Annales, Series Historia et Sociologia, 24, 3, 463–476.

**Wallace, Julian (2018):** Modelling Contemporary Gatekeeping: The Rise of Individuals, Algorithms and Platforms in Digital News Dissemination. Digital Journalism, 6, 3, 274–293.

**Wolfgang, J. David, Blackburn, Hayley & Stephen McConnell (2020):** Keepers of the Comments: How Comment Moderators Handle Audience Contributions. Newspaper Research Journal, 41, 4, 433–454.

**Wulczyn, Ellery, Thain, Nithum & Lucas Dixon (2017):** Ex Machina: Personal Attacks Seen at Scale. In: Proceedings of the 26th International Conference on World Wide Web. Perth, International World Wide Web Conferences Steering Committee. Geneva, International World Wide Web Conferences Steering Committee, 1391–1399.

**Wypych, Michał & Michał Bilewicz (2024):** Psychological Toll of Hate Speech: The Role of Acculturation Stress in the Effects of Exposure to Ethnic Slurs on Mental Health Among Ukrainian Immigrants in Poland. Cultural Diversity and Ethnic Minority Psychology, 30, 1, 35–44.

**Zuboff, Shoshana (2019):** The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power. New York, PublicAffairs.

**ZJRS (2004):** Zakon o javni rabi slovenščine. Uradni list RS, št. 86/04. https://www.pisrs.si/Pis.web/pregledPredpisa?id=ZAKO3924 (last access: 2025-11-28).

**ZMed (2006):** Zakon o medijih. Uradni list RS, št. 110/06. https://pisrs.si/Pis.web/pregledPredpisa?id=ZAKO1608 (last access: 2025-11-28).

**ZMed-1 (2025):** Zakon o medijih. Uradni list RS, št. 69/2025. https://www.uradni-list.si/glasilo-uradni-list-rs/vsebina/2025-01-2460 (last access: 2025-11-28).

**Žurnal24 (n.d.):** Pravno obvestilo. https://web.archive.org/web/20240418184034/https://www.zurnal24.si/pravno-obvestilo/ (last access: 2024-04-18).