Teka Komisji Prawniczej PAN Oddział w Lublinie
vol. XIX, 2026, no. 1, pp. 89-105
https://doi.org/10.32084/tkp.10636

HATE SPEECH IN THE PLATFORM AGE AT THE INTERSECTION OF INTERNATIONAL HUMAN RIGHTS STANDARDS, POLITICAL MOBILIZATION, AND LINGUISTIC MECHANISMS OF EXCLUSION

Dr. Bożena Iwanowska

Vizja University, Poland
e-mail: b.iwanowska@vizja.pl; http://orcid.org/0000-0003-1331-2866

Dr. habil. Yan Kapranov, University Professor

Vizja University, Poland
e-mail: y.kapranov@vizja.pl; http://orcid.org/0000-0003-2915-038X

Dr. Dawid Stadniczeńko

Vizja University, Poland
e-mail: d.stadniczenko@vizja.pl; http://orcid.org/0000-0001-6826-3366

Mariusz Dudek, MA

Vizja University, Poland
e-mail: m.dudek@vizja.pl; http://orcid.org/0000-0002-7316-0617

Abstract. Hate speech poses a growing challenge for democratic societies in the platform-mediated public sphere. This article develops an interdisciplinary, conceptual–analytical framework integrating international and European human rights law, political analyses of mobilization and institutional stress, and linguistic approaches to exclusionary meaning-making. Legally, it examines the regulation of hate speech through the balance between freedom of expression and the protection of dignity and equality, with particular attention to Article 20 of the International Covenant on Civil and Political Rights and Articles 10 and 17 of the European Convention on Human Rights, as interpreted by the European Court of Human Rights. Politically, hate speech is analyzed as a strategic resource in polarized competition that facilitates boundary-making and undermines civic trust. Linguistically, the article shows why hate speech cannot be identified through lexical markers alone, highlighting indirect and coded hostility, pragmatic speech acts, and dehumanizing metaphors. It argues that effective responses require multi-instrument governance combining proportionate legal enforcement, preventive measures, and accountable platform cooperation, and advances a multidimensional account of online hostility under platform conditions.

ISSN 1899-7694
e-ISSN 2719-7379

Keywords: freedom of expression; human rights adjudication; political polarization; critical discourse analysis; platform governance

INTRODUCTION

Hate speech has emerged as a central normative challenge for contemporary democracies, particularly within the platform-mediated public sphere. In its broadest sense, it encompasses expressions that spread, incite, promote, or justify hostility and discrimination against individuals or groups defined by protected or socially salient characteristics (such as ethnicity, religion, nationality, sexual orientation, or gender identity) and that may contribute to social exclusion, intimidation, or violence (Council of Europe, 1997; ECRI, 2016). The expansion of social media, combined with algorithmic models that privilege engagement and visibility, has intensified both the circulation and the persistence of such content, while simultaneously placing strain on traditional legal and institutional responses [Banks 2010; Florio et al. 2020]. Under these conditions, hate speech increasingly shapes the contours of public debate and participation, rather than remaining confined to isolated instances of individual harm.

Across disciplines, research has documented the multifaceted harms associated with hate speech. At the individual level, it undermines dignity, personal security, and psychological well-being; at the collective level, it reinforces stigma, normalizes marginalization, and contributes to climates of intimidation that discourage participation in public life [Gelber and McNamara 2015; Obrębska 2020]. Empirical studies further suggest that repeated exposure to hateful discourse, whether as a target, participant, or bystander, is linked to a greater likelihood of reproducing such communication, highlighting the role of social norms and peer environments in sustaining hostile speech practices [Wachs et al. 2021]. These dynamics indicate that hate speech functions not merely as isolated expression, but as a social practice capable of reproducing discriminatory hierarchies and shaping behavioral trajectories over time [Bera 2019; Wachs et al. 2021].

Against this background, hate speech cannot be adequately understood through a single analytical lens. Legal frameworks emphasize the tension between freedom of expression and the protection of dignity and equality; political analysis points to the role of exclusionary discourse in mobilization and institutional stress; linguistic research reveals the indirect, pragmatic, and metaphorical mechanisms through which hostility is normalized and circulated. Bringing these perspectives together allows for a more precise account of how hate speech operates under platform conditions and why responses limited to one dimension – whether doctrinal, political, or discursive – remain insufficient.

This article employs an interdisciplinary, conceptual–analytical approach that integrates three complementary perspectives: (1) a legal analysis of international and European human rights standards governing the limits of freedom of expression and the protection of dignity and equality – most notably Article 20 of the ICCPR and Articles 10 and 17 of the ECHR as interpreted in ECtHR case law through context-sensitive proportionality reasoning; (2) a political analysis that treats hate speech as a strategic resource of mobilization under conditions of polarization and institutional stress, intensified by platform architectures and mechanisms of algorithmic amplification; and (3) a linguistic and discourse-analytic perspective that explains how hostility and exclusion are produced and normalized through indirectness, pragmatic speech acts, dehumanizing metaphors, and coded ("dog-whistle") communication. The aim of the study is to develop an integrated analytical framework for a more precise account of hate speech in the platform-mediated public sphere and to derive governance-oriented implications that underscore the need for multi-instrument responses combining proportionate legal enforcement, preventive measures, and accountable cooperation with digital platforms. The research problem is centered on how hate speech should be conceptualized and analyzed under platform conditions in a manner that simultaneously captures the constraints of human rights adjudication, the political dynamics of mobilization and erosion of civic trust, and the linguistic mechanisms of indirect and coded hostility, thereby enabling the specification of adequate regulatory and institutional responses.

1. FREEDOM OF EXPRESSION AND ITS LIMITS: INTERNATIONAL AND EUROPEAN STANDARDS

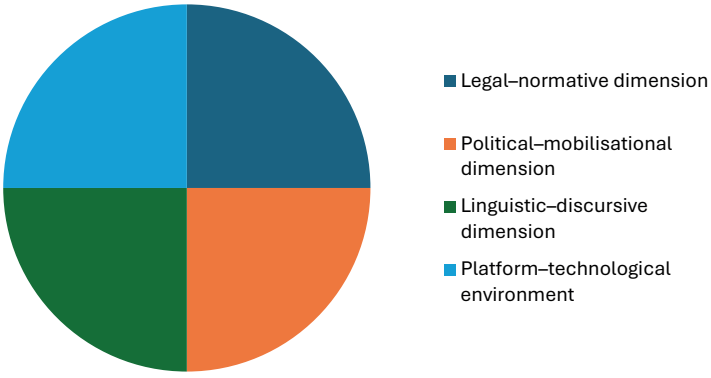
Hate speech is inseparable from the core democratic tension surrounding freedom of expression. Liberal constitutional orders protect speech as a foundation of pluralism, tolerance, and open debate, while simultaneously recognising that expression is not absolute when it undermines the rights and security of others. This tension is explicitly reflected in international and European human rights law. Article 20 of the International Covenant on Civil and Political Rights obliges States Parties to prohibit advocacy of national, racial, or religious hatred that constitutes incitement to discrimination, hostility, or violence (International Covenant on Civil and Political Rights [ICCPR], 1966/1977, Article 20). Within the European system, the European Convention on Human Rights guarantees freedom of expression while allowing restrictions that are "necessary in a democratic society" for legitimate aims, including the protection of the rights of others (European Convention on Human Rights [ECHR], 1950/1993, Article 10(2)).

The case law of the European Court of Human Rights reflects a dual approach to hate speech. In certain circumstances, expressions that negate the fundamental values of the Convention are excluded from protection under Article 17, whereas in others they are assessed under Article 10 through a contextual proportionality analysis (European Court of Human Rights, 2023). While the Court has consistently affirmed that freedom of expression extends to speech that "offends, shocks or disturbs," it has also accepted that democratic societies may penalize or prevent expressions that spread, incite, promote, or justify hatred based on intolerance, provided that any interference remains proportionate (*Handyside v. the United Kingdom; Erbakan v. Turkey*).

These standards reveal a persistent governance dilemma. Public authorities are required to protect vulnerable groups and prevent escalation into discrimination or violence, while avoiding regulatory overreach, politicized enforcement, or chilling effects that would undermine pluralism itself [Kapelańska-Pręgowska and Pucelj 2023; Bojanowski 2021]. This dilemma is intensified under platform conditions, where content circulates transnationally and moderation is increasingly exercised by private intermediaries operating under distinct incentive structures and accountability regimes [Banks 2010]. As a result, contemporary responses to hate speech increasingly rely on multi-instrument approaches that combine legal enforcement with preventive measures and cooperation with online platforms [Paz et al. 2020; Kapelańska-Pręgowska and Pucelj 2023].

Understanding how these legal constraints interact with political mobilization dynamics and linguistic mechanisms of exclusion is therefore essential for explaining how harmful communication is produced, normalized, and governed in the platform age.

Figure 1. Conceptual distribution of analytical dimensions of hate speech in the platform age.



Source: Author's own conceptual framework.

The figure illustrates analytically distinct yet interrelated dimensions of hate speech discussed in this article and does not represent empirical

measurement or proportional prevalence. It visualizes the integrated analytical perspective adopted here, linking legal regulation under international and European human rights law (ICCPR, 1966/1977; European Court of Human Rights, 2023), political dynamics of mobilization and institutional stress under conditions of polarization and platform amplification [Wodak 2021; Paz et al. 2020], and linguistic mechanisms through which hostility is produced, normalized, and circulated in discourse [van Dijk 1992; Culpeper 2011]. Together, these dimensions provide the basis for a governance-oriented synthesis focused on proportionality, prevention, and accountable cooperation with digital platforms [Kapelańska-Pręgowska and Pucelj 2023; Paz et al. 2020].

2. HATE SPEECH IN INTERNATIONAL AND EUROPEAN HUMAN RIGHTS LAW

2.1. Conceptualizing Hate Speech: Harm, Discrimination, and Social Impact

Hate speech is widely understood as forms of expression that promote or justify hostility and discrimination against specific social groups, and it has become a salient problem in the contemporary, platform-mediated public sphere. In a global context, the phenomenon has been examined across disciplines such as psychology, law, and information technology, reflecting its multidimensional character. The expansion of digital communication has further intensified concerns about the social effects of hate speech and the adequacy of existing regulatory responses.

Research consistently demonstrates that the harms of hate speech operate at both individual and collective levels. It undermines the dignity, safety, and psychological well-being of targeted individuals, while at the group level it reinforces stigma, normalizes marginalization, and contributes to climates of hostility and intimidation [Gelber and McNamara 2015; Obrębska 2020]. Studies focusing on young people indicate that exposure to hateful discourse, whether as witnesses or participants, increases the likelihood of reproducing such behavior, underscoring the role of social norms and peer environments in sustaining hostile communication [Wachs et al. 2021]. These dynamics support the view that hate speech functions not merely as isolated expression, but as a social practice capable of reproducing discriminatory hierarchies and patterns of aggression [Bera 2019].

From a broader social perspective, hate speech facilitates the marginalization and exclusion of minority groups, thereby implicating core human rights concerns. Attitudes toward the regulation of hate speech are themselves shaped by social and ideological orientations, with authoritarian dispositions tending to support stricter prohibitions, while hierarchical worldviews

are more tolerant of exclusionary expression [Bilewicz et al. 2015]. Within legal scholarship, this tension is reflected in debates over the scope and limits of freedom of expression. As Bojanowski observes, legal frameworks must align freedom of speech with standards aimed at protecting individuals and groups from hatred and violence, as freedom of expression cannot serve as a justification for discriminatory harm [Bojanowski 2021].

At the European level, the Council of Europe has adopted a broad definition of hate speech, encompassing the formulation, dissemination, and justification of hatred and discrimination in the public sphere against individuals, groups, and minorities, expressed through a wide range of communicative forms.[1]

2.2. The Dual Approach of the European Court of Human Rights: Articles 10 and 17 ECHR

Convention for the Protection of Human Rights and Fundamental Freedoms, done at Rome on 4 November 1950, as subsequently amended by Protocols Nos. 3, 5 and 8 and supplemented by Protocol No. 2[2] (hereinafter: the *European Convention on Human Rights*) provides for two approaches to hate speech. The first approach consists in excluding such expression from the protection of the Convention, pursuant to Article 17 (prohibition of abuse of rights), where the comments express hatred and negate the fundamental values of the Convention. The second approach consists in imposing limitations on Convention protection, pursuant to Article 10(2) (freedom of expression), applied where comments, although constituting hate speech, are not capable of destroying the fundamental values of the Convention (European Court of Human Rights, 2023).

Hate speech constitutes a serious challenge within the European system of human rights protection, particularly under the framework of the European Convention on Human Rights. Although Article 10 guarantees freedom of expression, this right is not absolute and may be subject to restrictions where expression undermines the rights, dignity, or security of others. As O'Flaherty observes, the growing prevalence of hate speech places increasing strain on the balance between freedom of expression and protection against discrimination within the European human rights regime [O'Flaherty 2017].

The case law of the European Court of Human Rights reflects this tension. The Court has consistently affirmed that "freedom of expression constitutes

[1] Appendix to Recommendation No. R (97) 20 of the Committee of Ministers of the Council of Europe, of 30 October 1997.

[2] *Convention for the Protection of Human Rights and Fundamental Freedoms, done at Rome on 4 November 1950, as subsequently amended by Protocols Nos. 3, 5 and 8 and supplemented by Protocol No. 2* (Journal of Laws of 1993, No. 61, item 284 as amended).

one of the essential foundations of a [democratic] society," extending not only to information and ideas that are favorably received, but also to those that "offend, shock or disturb," provided that any restrictions imposed are proportionate to the legitimate aim pursued (*Handyside v. the United Kingdom, para. 49*). At the same time, the Court has recognized that tolerance and respect for the equal dignity of all persons form the basis of a pluralistic democratic order, and that it may therefore be necessary to penalize or prevent expressions that spread, incite, promote, or justify hatred based on intolerance, subject to proportionality requirements (*Erbakan v. Turkey*, para. 56).

In legal scholarship, it is noted that the European Court of Human Rights has consistently emphasized that, while adjudication remains primarily within the competence of domestic courts, the imposition of a custodial sentence for a media-related offence, even if suspended, can only exceptionally be compatible with journalists' freedom of expression, particularly in cases involving hate speech or incitement to violence [Nowicki 2021] (Sallusti v. Italy, 2019, para. 59). The Court has further underlined that criminal sanctions against expressions of hatred may be justified only as a measure of last resort. Where conduct amounting to hate speech constitutes a serious attack on an individual's physical or psychological integrity, however, effective criminal-law responses may be required to ensure adequate protection and deterrence. This requirement has been affirmed in cases involving direct verbal assaults and threats motivated by discriminatory attitudes (Király and Dömötör v. Hungary, 2017, para. 76; Alković v. Montenegro, 2017, para. 8, 11, 65, 69; Beizaras and Levickas v. Lithuania, 2020, para. 111, 128).

2.3. Incitement, Context, and Proportionality in ECHR Case Law

In its case law, the European Court of Human Rights has clarified that incitement to hatred does not require an explicit call for violence or the commission of a specific criminal offence. Abusive treatment, ridicule, defamation of social groups, or incitement to discrimination may suffice to justify state intervention where such expressions undermine the dignity or security of targeted groups. Political statements driven by religious, ethnic, or cultural prejudice have therefore been recognized as posing risks to social peace and democratic stability (Dmitriyevskiy v. Russia, 2017, para. 99; Ibragim Ibragimov and Others v. Russia, 2018, para. 94).

The Court has further held that verbal attacks directed at an individual through the vilification of the group to which that person belongs may justify authorities prioritizing the protection of those affected over the speaker's reliance on freedom of expression (Carl Jóhann Lilliendahl v. Iceland, 2020). Particular weight is given to generalizing statements that portray entire ethnic, religious, or comparable groups in a negative or hostile manner,

especially where such statements contribute to the spread or justification of intolerance (Soulas and Others v. France, 2008, para. 40-43; Le Pen v. France, 2010; Norwood v. the United Kingdom, 2004; W. P. and Others v. Poland, 2004; Pavel Ivanov v. Russia, 2007; Féret v. Belgium, 2009, para. 71; Hizb ut-Tahrir and Others v. Germany, 2012, para. 73).

When assessing the proportionality of interference under Article 10, the Court consistently examines the broader context of the impugned statements, including their content, form, potential impact, and the social or political circumstances in which they were made. This contextual assessment requires consideration of whether the statements, viewed as a whole, can reasonably be understood as direct or indirect calls for violence, justification of violence, dissemination of hatred, or promotion of intolerance. Such evaluation is necessarily holistic and cannot be reduced to any single factor in isolation (Atamanchuk v. Russia, 2020, para. 50) [Mizerski 2021].

2.4. Institutional Actors and Soft Law: The Role of ECRI and International Standards

At the European level, a central role in addressing hate speech is played by the European Commission against Racism and Intolerance (ECRI), operating within the framework of the Council of Europe. Through its general policy recommendations and monitoring activities, ECRI seeks to counteract the spread of hatred and to promote tolerance and non-discrimination as core human rights principles. In its normative framework, ECRI defines hate speech broadly as "all forms of expression which spread, incite, promote or justify racial hatred, xenophobia, antisemitism or other forms of hatred based on intolerance, including intolerance expressed in the form of aggressive nationalism and ethnocentrism, discrimination and hostility against minorities."[3]

2.5. Hate Speech in Public International Law: Balancing Protection and Expression

In public international law, hate speech is addressed at the intersection of multiple protected interests, including the right to life, freedom from discrimination, and freedom of expression. International human rights norms therefore conceptualize hate speech not as an isolated category of expression,

[3] ECRI General Policy Recommendation No. 15 on combating hate speech, accessed on 15 November 2024: http://hudoc.ecri.coe.int/eng?i=REC-15-2016-015-ENG; ECRI General Policy Recommendation No. 7 on national legislation to combat racism and racial discrimination, accessed on 15 November 2024: http://hudoc.ecri.coe.int/eng?i=REC-07rev-2003-008-POL; ECRI General Policy Recommendation No. 6 on combating the dissemination of racist, xenophobic and antisemitic material via the Internet, accessed on 15 November 2024: http://hudoc.ecri.coe.int/eng?i=REC-06-2001-001-ENG [accessed: 11.02.2026].

but as a phenomenon requiring careful balancing between the prevention of violence and discrimination and the protection of expressive freedoms [De Fretes et al. 2023]. A central normative reference in this regard is the International Covenant on Civil and Political Rights, which obliges States Parties to prohibit "any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence" (ICCPR, 1966/1977, Article 20).

From an enforcement perspective, hate speech is increasingly analyzed in relation to its contribution to violence against minority groups. Empirical and policy-oriented research highlights a link between the circulation of hateful discourse and discriminatory or violent outcomes, reinforcing calls for effective legal regulation [Paz et al. 2020]. At the same time, the implementation of hate speech provisions remains uneven across jurisdictions, reflecting divergent legal traditions, political priorities, and levels of public awareness, which complicates transnational enforcement and cooperation [Pelor 2023]. In the global context, such divergences may generate tensions, particularly where hate speech is instrumentalized as a political tool [Iyiola 2024].

Digital communication technologies have further intensified these challenges. The rapid and transnational dissemination of content via online platforms has strained traditional regulatory mechanisms and exposed gaps in jurisdictional control and enforcement capacity [Banks 2010]. As a result, international scholarship and policy increasingly point toward multi-layered responses combining legal regulation with education, preventive strategies, and structured cooperation with online platforms [Kapelańska-Pręgowska and Pucelj 2023].

3. HATE SPEECH AS MOBILIZATION, BOUNDARY-MAKING, AND INSTITUTIONAL STRESS

Building on the international public law framework outlined above, political science explains how hate speech acquires political functionality within democratic systems. Beyond constituting a harmful form of expression, hate speech may operate as a strategic resource in political competition, facilitating group boundary-making and contributing to institutional stress through polarization and the erosion of civic trust [Wodak 2021; Gelber and McNamara 2015].

Exclusionary discourse is frequently deployed by political actors and networked publics to consolidate support by constructing emotionally charged distinctions between "us" and "them." Targeted groups are framed as threats to public order, cultural continuity, welfare systems, or national security – framings that gain particular traction under conditions of crisis or heightened social anxiety [Wodak 2021]. Linguistic constructions of citizenship

and belonging play a central role in this process, shaping political identity formation and enabling exclusionary narratives to gain social legitimacy [Iwanowska and Kapranov 2025]. As such frames become embedded in agenda-setting and securitization dynamics, democratic debate may shift from policy-oriented deliberation toward moralized conflicts over loyalty and belonging.

When routinised and politically rewarded, hate speech can function as a form of symbolic political violence. Rather than merely expressing hostility, it disciplines participation, normalizes social hierarchies, and signals unequal entitlement to dignity and voice [Gelber and McNamara 2015]. These dynamics intersect with broader processes of legitimacy construction, as ideological and cultural frameworks influence which forms of exclusion and coercion are perceived as acceptable or justified within a given political community [Iwanowska 2025a]. From this perspective, hate speech directly challenges democratic legitimacy, which rests on equal participation and mutual recognition.

Political science and prevention-oriented research further emphasize the contextual nature of escalation risks. The "dangerous speech" framework highlights how harm depends not solely on semantic content, but on speaker authority, audience susceptibility, crisis environments, and dissemination infrastructures that amplify repetition and normalization [Benesch 2014]. This helps explain why similar statements may remain marginal in one context yet become escalatory in another, particularly when they dehumanize targets or imply the necessity of exceptional measures.

Platform architectures intensify these dynamics. Social media systems reward engagement, accelerate diffusion, and foster homogeneous publics, enabling hate speech to function as a mobilization accelerator and lowering coordination costs for harassment or intimidation campaigns. Empirical research indicates that online hostility may translate into offline harm under identifiable conditions: anti-refugee sentiment on social media has been shown to predict hate crimes, while extremist violence can trigger measurable spikes in online hateful discourse, revealing feedback loops between offline shocks and online mobilization [Müller and Schwarz 2021; Olteanu et al. 2018].

From a governance perspective, these dynamics underscore the need for resilience-oriented policy responses. Overly broad regulation risks chilling legitimate dissent, while weak enforcement allows intimidation and exclusion to become normalized. Accordingly, political science points toward multi-instrument approaches that combine clearly defined legal thresholds, transparent and proportionate enforcement, and preventive measures such as education, media literacy, and counter-speech capacity [Benesch 2014; Paz et al. 2020]. In the platform environment, such strategies must also address algorithmic amplification and accountability, as trust in democratic institutions increasingly depends on the perceived legitimacy of digital governance arrangements [Banks 2010; Iwanowska 2025b].

4. HATE SPEECH AS DISCOURSE PRACTICE, PRAGMATIC ACTION, AND MEANING-MAKING IN CONTEXT

From a linguistic perspective, hate speech is best understood as a context-sensitive discourse practice rather than a fixed set of prohibited words. Linguistic analysis explains how hostility is produced through lexico-grammatical choices, pragmatic strategies, and genre-specific conventions (e.g., comments, memes, headlines, political slogans), and why exclusionary meanings may remain socially effective even in the absence of explicit slurs. This is particularly salient in platform environments, where brevity, virality, and multimodality (*text–image–emoji–hashtag*) favor condensed meanings and indirect signaling [van Dijk 1992; Reisigl and Wodak 2016].

4.1. Beyond slurs: indirectness, deniability, and the discursive normalization of hostility

Discourse-analytic research shows that hate speech often operates through indirect strategies such as insinuation, irony, humour, euphemism, and coded references. These forms allow speakers to activate exclusionary meanings within in-group audiences while preserving plausible deniability. A classic example is the use of disclaimers (e.g., "I'm not racist, but…"), which formally distance the speaker from prejudice while shifting responsibility onto the targeted group and reproducing inequality [van Dijk 1992]. Consequently, the identification of hate speech cannot rely on lexical inventories alone, but requires analysis of stance, implicature, and contextual uptake [Reisigl and Wodak 2016].

At the level of discourse structure, hostile communication frequently relies on recurrent argumentative topoi such as threat, contamination, betrayal, or burden. These topoi function as inferential shortcuts, enabling a transition from descriptive claims ("they are like this") to normative conclusions ("therefore exclusion is justified") without explicit incitement [Reisigl and Wodak 2016; Wodak 2021]. This helps explain how exclusionary discourse can exert political and legal pressure while remaining formally below criminal thresholds.

4.2. Pragmatics and speech acts: what hate speech does

From a pragmatic perspective, hate speech should be understood as social action rather than mere representation. Hostile utterances perform speech acts such as insulting, humiliating, threatening, excluding, or delegitimizing, and may produce coercive effects even without explicit calls to violence. These effects depend on illocutionary force (the act performed) and perlocutionary impact (the consequences triggered), both of which are shaped by speaker authority, audience alignment, and situational context [Butler 1997].

An additional interactional layer concerns linguistic aggression and impoliteness. Research on impoliteness identifies recurrent strategies of face attack – ridicule, name-calling, moral condemnation, or mock politeness – which are intensified in online environments by anonymity and attention incentives [Culpeper 2011]. Such practices sustain hostility through routine interaction rather than overt ideological declaration

4.3. Dehumanization, metaphor, and semantic framing

Linguistics provides tools for analyzing dehumanization as a mechanism of moral exclusion. As summarized in Table 1, hate speech frequently relies on metaphorical frames, such as animalization, disease, waste, invasion, or criminalization, that compress complex social identities into morally devalued categories. By reframing targets as risks, contaminants, or threats rather than moral agents, these frames reduce empathy and legitimize harsh or exclusionary treatment [Reisigl and Wodak 2016; Wodak 2021].

From a corpus-based perspective, such processes can be operationalized through analyses of semantic prosody and collocation patterns, including the repeated association of group labels with lexical fields of danger, impurity, or invasion [Reisigl and Wodak 2016]. In practice, these metaphors often appear in clusters, for example, linking disease imagery with invasion narratives, thereby intensifying perceptions of existential threat and lowering moral inhibitions against exclusionary responses [Wodak 2021].

Table 1. Dehumanising metaphors and semantic frames in hate speech discourse.

Type of semantic frame	**Typical metaphorical domain**	**Illustrative expressions**	**Discursive function**
Animalisation	Animals / vermin	*"rats", "cockroaches", "parasites", "vermin", "beasts", "pack animals", "hyenas", "apes", "insects"*	Reduces targets to instinct-driven beings; legitimises control, domination, or expulsion
Disease metaphors	Illness / infection	*"virus", "plague", "infection", "contagion", "disease", "epidemic", "cancer", "pathogen", "rot"*	Frames groups as threats to collective health; normalises exclusion, quarantine, or eradication
Waste metaphors	Dirt / pollution	*"filth", "trash", "human garbage", "pollution", "waste", "scum", "toxic elements", "dirt", "contamination"*	Constructs moral impurity; justifies removal, cleansing, or social exclusion
Invasion metaphors	War / natural disaster	*"flood", "invasion", "swarm", "tidal wave", "onslaught", "occupation", "siege", "overflow", "takeover"*	Depicts groups as overwhelming forces; enables securitization and emergency framing

Type of semantic frame	Typical metaphorical domain	Illustrative expressions	Discursive function
Criminalization frames	Crime / deviance	*"criminals", "rapists", "gangs", "thugs", "predators", "lawbreakers", "offenders", "delinquents"*	Associates identity with danger; legitimises surveillance, punishment, and coercive control
De-individualisation	Mass / anonymity	*"they", "those people", "the masses", "hordes", "crowds", "numbers", "faceless groups"*	Erases individuality; facilitates collective blame and moral disengagement

Source: Author's own conceptual framework, informed by critical discourse studies [van Dijk 1992; Reisigl and Wodak 2016; Wodak 2021].

4.4. Coded hostility and "dog-whistles": indexical meaning in platform publics

A central challenge for both research and regulation is coded hostility, often described as dog-whistle communication. Dog-whistles rely on indexical meaning: expressions that appear neutral to general audiences but activate shared hostile interpretations within specific communities. Because such meanings depend on intertextual knowledge and platform-specific repertoires (memes, slogans, hashtags), they frequently escape decontextualized screening mechanisms. Humour and irony further enhance deniability while preserving discriminatory force, complicating both moderation and legal assessment [Young 2025; van Dijk 1992].

4.5. Methodological contribution: corpus linguistics and the Discourse-Historical Approach

To complement the legal and political analyses, the linguistic perspective can be operationalized through two empirically oriented research designs. First, corpus-based annotation of platform content allows for systematic profiling of targets, speech acts, degrees of directness, dehumanization markers, and pragmatic devices, enabling comparative analysis across time periods or political events. Second, the Discourse-Historical Approach (DHA) facilitates longitudinal and intertextual tracing of how hostile labels and topoi circulate across actors, genres, and platforms, reconstructing the argumentative trajectory from representation to implied policy conclusions. This approach is particularly valuable for operationalizing the legal concept of context in proportionality assessments within human rights adjudication [Reisigl and Wodak 2016].

Taken together, these linguistic tools explain how hate speech is produced, normalized, and rendered resilient under formal condemnation, thereby

complementing the article's legal emphasis on contextuality and the political-science focus on mobilization and institutional stress (European Court of Human Rights, 2023) [Wodak 2021].

CONCLUSION AND FUTURE RESEARCH DIRECTIONS

Taken together, legal, political, and linguistic perspectives reveal hate speech not merely as prohibited expression, but as a communicative resource embedded in institutional arrangements and platform infrastructures. This article has argued that hate speech in the platform age should be understood not simply as unlawful or harmful expression, but as a form of communicative power embedded in legal norms, political competition, and linguistic practices shaped by digital infrastructures. By integrating international and European human rights law, theories of political mobilization and legitimacy, and discourse-analytic approaches to exclusion, the article has proposed a framework that conceptualizes hate speech as a relational phenomenon reproducing hierarchies of belonging, authority, and moral worth.

From a legal perspective, the analysis has shown that contemporary human rights regimes address hate speech primarily through context-sensitive balancing rather than categorical exclusion. The case law of the European Court of Human Rights demonstrates that proportionality, intent, audience impact, and socio-political context are central to legal assessment, bringing judicial reasoning closer to insights developed in discourse analysis and political theory. Law thus functions not only as a reactive mechanism, but also as a site where the boundaries of legitimate public discourse are continuously negotiated.

Politically, hate speech has been shown to operate as a resource of mobilization and legitimation, particularly under conditions of polarization and institutional stress. Exclusionary discourse contributes to the symbolic ordering of political communities by defining who is entitled to recognition and protection, and who may be constructed as a threat. Its effectiveness lies in its capacity to normalize exclusion below the threshold of explicit violence, thereby reshaping perceptions of what forms of inequality or coercion appear acceptable.

From a linguistic standpoint, the article has demonstrated that hate speech functions through indirectness, metaphor, pragmatic action, and indexical meaning. Dehumanizing frames, coded hostility, and strategic ambiguity allow exclusionary meanings to circulate while preserving deniability, complicating both legal adjudication and platform moderation. This confirms that the social harms of hate speech cannot be captured at the level of lexical content alone, but require attention to pragmatic force, repetition, and discursive normalization.

Taken together, these findings support a resilience-oriented approach to democratic governance in which the regulation of hate speech is treated as an infrastructural challenge rather than a purely doctrinal one. While legal sanctions remain necessary for the most severe forms of incitement, they are insufficient in isolation. Effective responses depend on the interaction of legal standards, political accountability, communicative norms, and platform design. In this context, algorithmic curation and amplification play a critical role in shaping the visibility and perceived legitimacy of hostile discourse, making questions of transparency and accountability central to democratic resilience.

Although this article has been conceptual in scope, it points to several directions for future research. Comparative discourse studies could examine how dehumanizing metaphors and coded hostility vary across languages and political cultures. Political communication research could analyze the relationship between hate-driven narratives, legitimacy construction, and institutional trust during periods of crisis. Finally, interdisciplinary governance research could assess how different models of platform regulation and algorithmic accountability influence both the circulation of hostile discourse and the legitimacy of regulatory interventions. Accordingly, the article provides a concrete answer to the research problem. Hate speech in the platform age should not be conceptualized as a fixed category identifiable by keyword lists, but as a relational, context-dependent social practice whose assessment requires the simultaneous consideration of: (1) the human-rights framework and its proportionality- and context-based model of qualification (including the distinction between exclusion from protection under Article 17 ECHR and permissible restrictions under Article 10(2) ECHR), (2) the political dynamics of mobilization and the erosion of civic trust under conditions of polarization, and (3) the discursive mechanisms through which hostility operates indirectly and indexically (e.g., via dehumanizing frames and dog-whistle communication) and is subsequently normalized and amplified within platform infrastructures. On this basis, the article derives a governance-oriented conclusion that adequate regulatory and institutional responses must be multi-instrumental. Beyond proportionate legal enforcement for the most severe forms of incitement, they should include preventive measures (e.g., education, media literacy, and counter-speech capacity) and accountable cooperation with platforms oriented toward transparency and the mitigation of amplification mechanisms that sustain hostile content.

In sum, hate speech in the platform age should be analyzed as a dynamic social practice embedded in legal, political, and communicative infrastructures. The framework developed here provides a basis for further interdisciplinary inquiry and for governance strategies aimed not only at limiting harm, but at strengthening the normative and institutional foundations of democratic life.

REFERENCES

Banks, James. 2010. "Regulating hate speech online." *International Review of Law, Computers & Technology* 24(3):233-39. https://doi.org/10.1080/13600869.2010.522323

Benesch, Susan. 2014. *Countering dangerous speech: New ideas for genocide prevention* (Working paper). United States Holocaust Memorial Museum.

Bera, Ryszard W. 2019. "Mowa nienawiści źródłem przemocy i agresji." *Annales Universitatis Mariae Curie-Skłodowska. Sectio J, Paedagogia-Psychologia* 32(3):59-66. https://doi.org/10.17951/j.2019.32.3.59-66

Bilewicz, Michał, et al. 2015. "When authoritarians confront prejudice. Differential effects of SDO and RWA on support for hate-speech prohibition." *Political Psychology* 38(1):87-99. https://doi.org/10.1111/pops.12313

Bojanowski, Tomasz. 2021. "*Wybrane prawnokarne aspekty mowy nienawiści w kontekście standardów ochrony wolności słowa*." *Prawo w Działaniu* 47:168-86. https://doi.org/10.32041/pwd.4710

Butler, Judith. 1997. *Excitable speech: A politics of the performative*. Routledge.

Culpeper, Jonathan. 2011. *Impoliteness: Using language to cause offence*. Cambridge University Press.

De Fretes, Diego Romario, et al. 2023. "Challenges in enforcing hate speech laws in Indonesian politics." *International Journal of Humanities, Social Sciences and Business (INJOSS)* 2(3):418-42. https://doi.org/10.54443/injoss.v2i3.89

Florio, Komal, et al. 2020. "Time of your hate: the challenge of time in hate speech detection on social media." *Applied Sciences* 10(12):4180. https://doi.org/10.3390/app10124180

Gelber, Katharine, and Luke McNamara. 2015. "Evidencing the harms of hate speech." *Social Identities* 22(3):324-41. https://doi.org/10.1080/13504630.2015.1128810

Iwanowska, Bożena, and Yan Kapranov. 2025. "Homo politicus and res publica today: Linguistic reappraisal with student survey evidence." *Horyzonty Polityki* 16(56):271-90. https://doi.org/10.35765/HP.2732

Iwanowska, Bożena. 2025a. "Algorithmic legitimacy and the digital state: Rethinking governance, trust, and accountability in the age of AI." *AI, Law, Politics: An Interdisciplinary Journal on Human–AI Interaction in Legal and Political Systems* 1(2):130-49. https://doi.org/10.5709/alp-01.02.2025-03

Iwanowska, Bożena. 2025b. "Cross-cultural perceptions into the ideological foundations of power legitimacy: Empirical evidence from international students." *Acta Humanitatis* 3(1):4-24. https://doi.org/10.5709/ah-03.01.2025-01

Iyiola, Damilola E. 2024. *The International Human Rights Law and Social Media Regulation* [Preprint]. https://doi.org/10.31219/osf.io/hfcey

Kapelańska-Pręgowska, Julia, and Maja Pucelj. 2023. "*Freedom of expression and hate speech: Human rights standards and their application in Poland and Slovenia*." *Laws* 12(4):64. https://doi.org/10.3390/laws12040064

Mizerski, Rafał. 2021. "Glosa do wyroku Europejskiego Trybunału Praw Człowieka z dnia 28 sierpnia 2018 r. w sprawie *Savva Terentyev p. Rosji* (10692/09)." *Studia Prawnoustrojowe* 51:215-26. https://doi.org/10.31648/sp.6405

Müller, Karsten, and Carlo Schwarz. 2021. "*Fanning the flames of hate: Social media and hate crime.*" *Journal of the European Economic Association* 19(4):2131-167. https://doi.org/10.1093/jeea/jvaa045

Nowicki, Marek A. 2021. "Komentarz do art. 10 Europejskiej Konwencji Praw Człowieka." In *Wokół Konwencji Europejskiej. Komentarz do Europejskiej Konwencji Praw Człowieka* (8th ed.). Wolters Kluwer.

O'Flaherty, Michael. 2017. "Ochrona praw człowieka w dzisiejszej Europie." *Ruch Prawniczy, Ekonomiczny i Socjologiczny* 79(1):49-57. https://doi.org/10.14746/rpeis.2017.79.1.4

Obrębska, Monika. 2020. "*Contempt speech and hate speech: Characteristics, determinants and consequences.*" *Annales Universitatis Mariae Curie-Skłodowska. Sectio J, Paedagogia-Psychologia* 33(3):9-20. https://doi.org/10.17951/j.2020.33.3.9-20

Olteanu, Alexandra, et al. 2018. "*The effect of extremist violence on hateful speech online.*" *Proceedings of the International AAAI Conference on Web and Social Media* 12(1). https://doi.org/10.1609/icwsm.v12i1.15040

Paz, María A., et al. 2020. "*Hate speech: A systematized review.*" *SAGE Open* 10(4). https://doi.org/10.1177/2158244020973022

Pelor, Stephanus. 2023. "*Law enforcement of hate speech criminals through social media based on Indonesia's positive law.*" *International Journal of Multidisciplinary Research and Literature* 2(3):347-58. https://doi.org/10.53067/ijomral.v2i3.123

Reisigl, Martin, and Ruth E. Wodak. 2016. "The discourse-historical approach." In *Methods of critical discourse studies, edited by Ruth* Wodak, and Michael Meyer, 3rd ed., 23-61. SAGE.

van Dijk, Teun A. 1992. "Discourse and the denial of racism." *Discourse & Society* 3(1):87-118. https://doi.org/10.1177/0957926592003001005

Wachs, Sebastian, et al. 2021. "*Playing by the rules? An investigation of the relationship between social norms and adolescents' hate speech perpetration in schools.*" *Journal of Interpersonal Violence* 37(21-22), NP21143–NP21164. https://doi.org/10.1177/08862605211056032

Wodak, Ruth. 2021. *The politics of fear: The shameless normalization of far-right discourse* (2nd ed.). SAGE.

Young, Jennifer. 2025. "Dogwhistles, discrimination, humour and the law: Regulating implicit messaging." *Open Library of Humanities* 11(2). https://doi.org/10.16995/olh.19789