HATE SPEECH

Contemporary issues and debates

HDV PUBLICATIONS

HRANT DINK FOUNDATION

Hrant Dink Foundation was established after the assassination of Hrant Dink in front of his newspaper Agos on January 19, 2007, in order to avoid similar pains and to continue Hrant Dink's legacy, his language and heart and his dream of a world that is more free and just. Democracy and human rights for everyone regardless of their ethnic, religious or cultural origin or gender is the Foundation's main principle.

The Foundation works for a Turkey and a world where freedom of expression is limitless and all differences are allowed, lived, appreciated, multiplied and conscience outweighs the way we look at today and the past. As the Hrant Dink Foundation 'our cause worth living' is a future where a culture of dialogue, peace and empathy prevails.

HATE SPEECH: CONTEMPORARY ISSUES AND DEBATES

ISBN 978-625-97168-0-0

DISCOURSE
DIALOGUE
DEMOCRACY
LABORATORY

editors
İdil Engindeniz, Zeynep Arslan

proofreader
Kate Ferguson

publication coordinators
Başak Can, Elif Erol, Buket Kapısız, Yasemin Korkmaz,
Pelin Önal, Tuğba Özsoy, Elif Yararbaş

design
Yasemen Cemre Gürbüz

graphic application
Selin Uluer

printed by
Sena Ofset Ambalaj Sanayi ve Ticaret. Ltd. Şti.
Yakuplu Mh. 194. Sk. 3.Matbaacılar Sit. N:1 D:465
Beylikdüzü İstanbul/Türkiye
T: (212) 613 38 46

Istanbul, March 2025

HDV PUBLICATIONS

Anarad Hığutyun Binası Papa Roncalli Sokak No: 128
Harbiye, 34373 Şişli, İstanbul
T: 0212 240 33 61
F: 0212 240 33 94
info@hrantdink.org
www.hrantdink.org

Utilizing Digital Technology for Social Cohesion, Positive Messaging and Peace by Boosting Collaboration, Exchange and Solidarity project is supported by the European Union and the Friedrich Naumann Foundation. The Hrant Dink Foundation is solely responsible for the content in the publication, which does not reflect the views of the supporters.

FRIEDRICH NAUMANN FOUNDATION For Freedom. Türkiye

HRANT DİNK VAKFI
HRANT DINK FOUNDATION
ՀՐԱՆԴ ՏԻՆՔ ՀԻՄՆԱԴՐԱՄ

Sabancı Üniversitesi

HATE SPEECH

Contemporary issues and debates

HDV
PUBLICATIONS

CONTENTS

Foreword

This report is the culmination of a year-long collaborative effort aimed at addressing the growing challenges of hate speech, disinformation and harmful discourse in digital spaces. This initiative was born out of the Hrant Dink Foundation's project on hate speech, **Utilizing Digital Technology for Social Cohesion, Positive Messaging and Peace by Boosting Collaboration, Exchange and Solidarity**[1], funded by the European Union and co-funded by the Friedrich Naumann Foundation. The project, carried out in partnership with Boğaziçi University and Sabancı University, aims to instigate an interdisciplinary effort to combat hate speech and discrimination in digital spaces.

This report is part of the ongoing work of the Hrant Dink Foundation, which was established in 2007 following the assassination of Hrant Dink, a prominent Armenian journalist and advocate for dialogue and understanding. Hrant Dink was the target of hate speech through the media, which played a significant role in the events leading to his assassination. The Foundation was established to continue his legacy by fostering dialogue, promoting peace, and addressing the societal issues that contribute to discrimination. Guided by these principles, the Foundation has been working to combat discrimination and foster coexistence through various initiatives, including a focus on hate speech and discriminatory discourse.

Since 2009, the **Media Watch on Hate Speech**[2] project has been a cornerstone of the Foundation's efforts to address hate speech and discrimination in Turkey. The aim of this project is to contribute to the civilian oversight of Turkey's print press, drawing attention to discriminatory and marginalizing language directed at various identities and groups. Through the systematic monitoring and reporting of national and local press, the project seeks to raise awareness and encourage inclusive discourse.

[1] **Hrant Dink Foundation.** (2022). *Utilizing digital technology for social cohesion, positive messaging and peace by boosting collaboration, exchange and solidarity*. Hrant Dink Foundation. https://hrantdink.org/en/asulis/activities/projects/utilizing-digital-technology-for-social-cohesion-positive-messaging-and-peace-by-boosting-collaboration-exchange-and-solidarity

[2] **Hrant Dink Foundation.** *(2016). Media watch on hate speech*. Hrant Dink Foundation. https://hrantdink.org/en/asulis/activities/projects/media-watch-on-hate-speech

Leveraging the expertise developed through the Media Watch on Hate Speech project, in 2016 the Foundation established the **ASULIS Discourse, Dialogue, Democracy Laboratory**[3], the first research center in Turkey dedicated to discourse studies. ASULIS, the name of which is derived from the Armenian verbs *asel* (to say) and *lsel* (to listen). ASULIS is a social sciences laboratory committed to combating discrimination, producing research on discourse, and supporting work in this field. It has expanded the scope of the Foundation's work, creating an interdisciplinary space for research, discussion and action on hate speech and discriminatory discourse.

In 2021, the **Hate Speech Digital Archive**[4] was launched. This publicly accessible digital archive encompasses more than a decade of systematic monitoring of hate speech in Turkish print media, conducted since 2009. Featuring thousands of categorized examples, the archive provides researchers, activists, and the general public with a comprehensive resource for understanding the patterns and impact of hate speech in the media.

Recognizing the increasing impact of digital platforms on public discourse, in 2021 the Foundation expanded its efforts to address hate speech and disinformation in online spaces. Through the project "Utilizing Digital Technology for Social Cohesion, Positive Messaging, and Peace by Boosting Collaboration, Exchange and Solidarity", the Foundation aimed to explore effective methods for identifying and addressing harmful narratives online.

Over the course of the project, 20 experts – comprised of academics, researchers, civil society representatives and activists from Turkey and abroad - came together to form a network of professionals from diverse fields such as discourse studies, computer science, linguistics, and civil society. While interdisciplinary in nature, the group maintained a shared focus on hate speech, with all members bringing relevant expertise and insights to the discussions. Between March 2023 and February 2024, the network met regularly in a series of thematic meetings to discuss various issues such as e-literacy, hate speech and disinformation on social media, digital activism, and innovative tools to combat harmful speech. Each meeting served as a space for collaboration, experience-sharing and knowledge exchange, informed by both local and global perspectives.

[3] **Hrant Dink Foundation.** (2016) *ASULIS discourse, dialogue, democracy laboratory*. Hrant Dink Foundation. https://hrantdink.org/en/asulis

[4] **Hrant Dink Foundation.** (2021) *Hrant Dink Foundation Archive*. Hrant Dink Foundation. https://archive.hrantdink.org/?l=en

The purpose of this report is to document the discussions and insights that emerged from these meetings. It is intended to serve as a comprehensive resource for anyone interested in understanding and addressing the challenges of hate speech and disinformation in today's digital age. This publication not only features contributions from network members but also aims to disseminate the network's learnings to a broader audience, serving as a resource for activists, civil society organizations, researchers, and policy-makers. The articles included in this report provide expert analysis and recommendations, functioning as a blueprint for tackling pressing issues.

The first part of the report examines the identification and categorization of hate speech, offering diverse definitions and insights into related concepts such as discrimination, dangerous speech and offensive language, while delving into the intersectional identities of hate speech targets. Subsequent sections examine policies surrounding hate speech, automated detection models, and strategies for countering hate speech beyond regulations. Contributions from experts across various disciplines enrich the report, creating an interdisciplinary dialogue among experts, complemented by editorial textboxes highlighting key moments of discussion and providing additional context for the reader. The format of this report is a reflection of the collaborative and multifaceted nature of this initiative.

We thank the experts in the network, the authors of this report, for their lively discussions and invaluable contributions to this field, and hope that this publication will multiply the ongoing efforts to promote social cohesion, peace and positive discourse in the digital realm.

IDENTIFICATION AND CATEGORIZATION OF HATE SPEECH AND OTHER RELATED CONCEPTS

Network members initiated a discussion on methods for preventing hate speech, starting by looking at the definition of the term. Regardless of the prevention methodology used, whether manual or through artificial intelligence, in order to identify and undertake the requisite work to eliminate and prevent them, it is crucial to first define hate speech and discriminatory discourses. Ongoing discussions in the field on the definition of the concept will form the basis for a more robust discussion on detection, prevention and intervention in the following sections of the report.

In defining the concept of hate speech, Yasemin İnceoğlu provided a summary of the limitations of the definition and the ongoing debates around it.

Different definitions of hate speech

Yasemin İnceoğlu

To begin, it is essential to dissect the term "hate speech" by examining its two constituent words: "hate" and "speech".

Some researchers attempt to elucidate "hate" as an emotion, creation, syndrome or even – taking a psychoanalytic approach – a spiritual abnormality. However, in the context of this document and of the other work realized by the Hrant Dink Foundation, it is considered as the result of an entire (political/social) system or approach.

When it comes to "speech", in linguistics, the concept of discourse often covers patterns of speech and language use, encompassing dialects and socially accepted expressions within a particular community. Discourse serves as an ideology rooted in social origins and encoded within language. Critical discourse analyst Van Dijk (2008) contends that controlling or generating discourse is imperative to exert mental control over society. However, it is important to recognize that discourse is influenced by subjective and psychological contexts, including the individuals involved, their intentions, and situational factors. Van Dijk (2008, pp. 107-108) underscores that the foremost prerequisite for controlling discourse is the management of its contextual elements.

Although "hate speech" finds widespread usage in legal, policymaking and academic circles, there is no universally accepted definition of the term within the framework of international human rights law or in national law. The concept remains a subject of ongoing discussion, particularly in its intersection with fundamental principles such as freedom of opinion and expression, non-discrimination and equality.

In a general sense, hate speech can be understood as any mode of expression wherein the speaker defames, demeans, or encourages hatred towards a particular group or category of individuals based on factors such as race, religion, skin color, sexual identity, gender identity, ethnicity, disability, or national origin. It is important here to underline that this list of categories changes and expands over time. Below is an overview, in chronological order, of some international documents illustrating the basics of the concept (even though their intentions were not to do so).

While Article 20/2 of the International Covenant on Civil and Political Rights (ICCPR)[5] does not provide a direct definition of hate speech, it has come to be considered a fundamental principle regarding the concept. The Article states that, "Any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence shall be prohibited by law." Adopted in 1966, this provision from the ICCPR covers a much narrower scope than more recent definitions of hate speech, defining hatred only on the basis of religion, race or nationality.

The Council of Europe (CoE) issued a recommendation regarding hate speech, defining it as encompassing, "all forms of expression which spread, incite, promote or justify racial hatred, xenophobia, anti-Semitism or other forms of hatred based on intolerance, including: intolerance expressed by aggressive nationalism and ethnocentrism, discrimination and hostility against minorities, migrants and people of immigrant origin."(1997)

Prepared through regional expert workshops organized by the Office of the United Nations High Commissioner for Human Rights (OHCHR), the 2012 Rabat Plan of Action included conclusions and recommendations adopted by experts. The "Rabat Plan of Action on the prohibition of advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence" indicates that, "to assess the severity of the hatred, possible elements may include the cruelty or intent of the statement or harm advocated, the frequency, quantity and extent of the communication." (OHCHR, 2012). Furthermore, a six-part threshold test was proposed for expressions considered criminal offences, which included the following items:

- Context
- Speaker
- Intent
- Content and form

[5] Adopted and opened for signature, ratification and accession by General Assembly resolution 2200A (XXI) of 16 December 1966 entry into force 23 March 1976. **Republic of Türkiye Ministry of Foreign Affairs**. (2024). *International Covenant on Civil and Political Rights*. Republic of Türkiye Ministry of Foreign Affairs. https://www.mfa.gov.tr/international-covenant-on-civil-and-political-rights.en.mfa

- Extent of the speech act
- Likelihood, including imminence

A more recent definition was offered by the United Nations (UN) in its 2019 “Strategy and Plan of Action on Hate Speech”. In this document, the term hate speech is understood as any kind of communication in speech, writing or behavior that attacks or uses pejorative or discriminatory language with reference to a person or group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, color, descent, gender or other element of their identity (UN, 2019). This definition is very broad in two senses: firstly, it refers to a wide variety of groups (by including “or other identity factor”), and secondly, the term “pejorative” covers a large range of discourse.

As we will observe throughout this document, beyond these attempts to reach a universally accepted definition of hate speech, social media companies, non-governmental organizations and researchers may have varying definitions of hate speech. However, before coming up with different definitions or talking about new concepts emerging in this area to fill the gaps left by the concept of hate speech, we would like to touch upon one of the most important debates surrounding the issue: the relationship between speech and action. Yasemin İnceoğlu opens up a space for this discussion with these questions:

“Is discourse merely an exchange of words, with those uttering them only accountable for their individual actions? Or should some discourse be regarded as a propagandistic element that fuels discrimination and ultimately legitimizes violence?”

In answer to these questions, in the following article Ayşecan Terzioğlu explains the relationship between the act of discrimination and hate speech.

Dangerous speech:
A framework for distinguishing harmful rhetoric

Susan Benesch

There are many competing definitions for hate speech, as well as for related terms such as toxic or offensive speech. This is in no small part because many of the harms included in definitions – such as insult or disrespect - are subjective and variable between individuals and cultures. Moreover, insult and disrespect can overlap or be confused with forms of speech that should be permitted in public discourse, such as criticism. It is therefore challenging to find consensus on which content constitutes hate speech in the margins of the category, especially between cultures and normative groups.

A related category of rhetoric can be identified more consistently and with less disagreement: dangerous speech, or content that incites people to condone or even commit violence against members of any other group. There are striking similarities or rhetorical patterns in dangerous speech, referred to as "hallmarks". Those hallmarks of dangerous speech include dehumanization, assertions of attack against women and girls of the in-group, describing an ostensible threat to the purity or integrity of the in-group, and "accusation in a mirror," which refers to telling members of an in-group that an out-group is planning to attack them, when the speaker plans to attack the out-group instead, and wants to convince the audience that such an attack would be defensive, in order to make it seem justified and even virtuous.

Another feature of nearly all dangerous speech is that it lowers normal social barriers against violence by convincing people to fear another group, and especially to perceive it or its members as a serious threat. Fear is likely to be a more powerful and relevant force in driving groups of people apart to the point of violence, than hatred (Leader Maynard & Benesch, 2018).

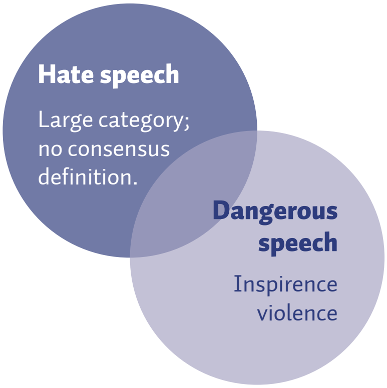
Since dangerous speech is identified using an analytical framework that calls on the analyst to make a prediction about the behavior of *other* people, the definition is not as subjective as that of hate speech.

Finally, the categorical boundaries of dangerous speech are not formed by including or excluding any types of identity group. This avoids a major inconsistency among definitions of hate speech, most of which do specify certain groups. As noted above, recent hate speech definitions include a wider variety of groups than older definitions, which were limited to ethnicity, race, religion and nationality, in keeping with the human rights law instruments written in the aftermath of World War II. Newer definitions, including those used by technology companies for content moderation, encompass such identity markers as caste, immigrant status, and gender identity. They generally exclude gender itself, however, as well as political opinion/affiliation and employment. This means that, for example, vitriolic attacks on journalists, though they inspire hatred and discrimination, generally do not count as hate speech. However, they certainly can, and often do, constitute dangerous speech.

Dangerous speech is distinct from hate speech since not all hate speech increases the risk of intergroup violence, and not all dangerous speech is hateful. To explain further: as Gelber (2021) argues in her analysis of systemic discrimination, hate speech may always increase some harms such as discrimination, however it cannot be said to increase violence in all cases, since some (though not enough) audiences are strongly disinclined to commit or condone violence. On the other hand, it is possible to convince people to condone violence by persuading them not to hate but to *fear* another group. In sum, hate speech and dangerous speech overlap in a Venn diagram as illustrated in figure 1.

Dangerous speech is commonly false – not surprising, since it describes whole groups of human beings in appalling terms. Unfortunately, people can be quite easily persuaded of misinformation (false assertions) or disinformation (false as-

Figure 1: Overlap of Hate Speech and Dangerous Speech



sertions that are spread knowingly or intentionally). And when falsehoods are frightening, people are more likely to spread them, even when they are not sure whether they are true. In such circumstances, people readily accept exaggerated or false messages (Leader Maynard & Benesch, 2018).

Scholars and speech regulators, especially at technology companies, have found the concept of dangerous speech useful for several reasons. First, as noted above, it is defined consequentially and objectively by its capacity to increase the risk of intergroup violence. Second, there is strong consensus in most societies against mass violence and on the value of preventing it. Third, there is no need to argue over which identity groups count, since any group can be targeted by dangerous speech. In studies of dangerous speech in many countries and historical periods, we have found examples aimed at all too many groups and types of group.

The relationship between discrimination and hate speech

Ayşecan Terzioğlu

Hate speech and discrimination are closely related with each other, to the extent that political scientist Katharine Gelber points out that hate speech can be understood as a discursive act of discrimination, which often harms the principles of equal opportunities and rights for all concerned (Gelber, 2021). Gelber elaborates that both hate speech and discrimination have the capacity to harm people directly or indirectly, depending on the political, economic and social context. This contextualization also has a temporal element, which includes the repetition of hate speech and discrimination over time, either separately or accompanying each other. Discrimination and hate speech can also turn into each other, and produce similar discourses and practices that target specific groups. However, discrimination is a much larger concept, since it is not only discursive, but also actional and institutional, expressed and implemented non-verbally.

Both hate speech and discrimination can often be based on race/ethnicity, class and gender, separately or in combination with each other in an intersectional way (Crenshaw, 2005). However, other demographic and socio-economic factors – such as people's age, marital status, occupation, level of education , political opinions, religious belief, and where they live – can also make them targets of hate speech and/or discrimination in particular contexts. Discrimination can be seen in individuals' actions, as in the case of a survey conducted in Germany looking at respondents' preferences for carpooling offers, where the *perceived* ethnic background of the driver was found to be an important criterion in the respondents' choices, based on prejudices against certain groups of immigrants in Germany, such as Turks and Italians (Liebe & Beyer, 2021).

To illustrate how acts of discrimination and hate speech surface, we can give the typical examples of people who are denied education, housing, health care or a job because of their gender, ethnicity, race or citizenship, and how hate speech is used as an "excuse" for this denial, as in the case of "they do not deserve these rights because they are (...)" In this case, hate speech is used to "legitimize" discrimination, and if this becomes a systematically repeated pattern, there is a risk of the formation of a vicious circle between hate speech and acts of discrimination. As such, interested parties, such as scholars, policymakers and content creators, should be especially wary of such patterns, which risk normalizing hate speech in particular societies or social contexts. Scholar of the philosophy of language and epistemology, Kindermann (2023) suggests that not all forms of discriminatory speech are morally sanctioned and legally regulated, even if they harm people in terms of excluding them from certain groups and deteriorating their interactions with people within those groups. Kindermann argues that it is difficult to define a lower threshold as "sufficiently harmful" for sanctioning and regulating, since supposedly "minor" and strong acts of discrimination share the same fundamental nature, and are part of the same continuum.

Working in the field of informatics, computing engineering and social psychology, Fortuna and Nunes (2018) claim that even subtle acts of discrimination, including jokes, should be considered as hate speech, since they are based on stereotypical generalizations and negative judgements and can have negative psychological effects on people. Discrimination can go beyond individuals' everyday life actions and become institutional, when they turn into patterns that shape institutional policies and procedures. For instance, certain companies' hiring and promoting patterns may show discrimination based on gender and/or sexual orientation, to the extent that women and LGBTI+ individuals have to struggle against the glass ceiling in order to be hired and/or promoted (Manzi & Heilman, 2021).[6]

[6] For a detailed discussion on hate speech and its results on daily life in the context of LGBTI+ migrants, the reader is invited to read the section of this report authored by Eser Selen.

Offensive language: Variability and challenges

Tommaso Caselli

Defining *offensive language* is not a trivial task. “Offence […] requires people to be offended” (O’Driscoll, 2020, p. 11): this opens up numerous questions concerning which words (if any) are offensive, how and when offense is triggered (ie., the situation in which a specific message is uttered and received by the participants, including their cultural background and experience).

Here, we will focus on the effects that words and sentences may have on the participants in a hypothetical discursive situation, rather than on the intentions of what has been said. Following O’Driscoll (2020), offensive language is seen as a scalar phenomenon: one can follow an imaginary continuum of the perceived offensiveness of what is said, and this scale is not objective. Offense is primarily a subjective phenomenon: the offense relies on the perception of the receiver.

A unique and shared definition of offensive language is missing, despite the attention that different disciplines (linguistics, media studies, communication studies, natural language processing, philosophy, psychology, law) have paid to this phenomenon. In Table 1, we illustrate the result of a non-systematic review conducted on Google Scholar[7] with some of the most common expressions used to characterize this area of research. The frequency refers to the number of entries retrieved by the search engine when fed with these expressions.

Table 1. Results from Google Scholars of the frequency of some expressions commonly used when referring to “offensive language”.

Term	Google scholar frequency
Toxic language	3.720.000
Abusive language	2.180.000
Offensive language	2.090.000
Taboo language	1.500.000
Obscene language	624.000
Insulting language	522.000
Swearing	220.000

[7] Search conducted on February 11, 2024.

Quite surprisingly, the most common term is "toxic language", while the expression "offensive language" occurs only in third position. A first remark, by exploring some of the descriptions in the text snippets associated with the returned results for each of these terms, is that they are used interchangeably, a behavior also observed by O'Driscoll (2020). This is not only an issue of preference for one expression over another but also an indication of relatively widespread disagreement among scholars on what exactly counts as offensive language. If we start comparing some of the definitions used to carve out the meaning and the nature of offensive language (see Table 2), we can observe commonalities and a general pattern indicating that offensive language is a very broad and general phenomenon that functions as a large umbrella term for more specific and heinous ones (e.g. hate speech).

Although there is a disagreement on the surface term used (toxicity vs. offensive language vs. obscene), all the definitions in Table 2 insist on the negative effects of the utterances on the receiver, which is at the core of offensive language.

Table 2. Comparable definitions of "offensive language".

Definition	**Source**
(…) the term "toxicity," defined as "a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion."	Perspective API[8]
(…) any word or string of words which has or can have a negative impact on the sense of self and /or wellbeing of those who encounter it - that is, it makes or can make them feel, mildly or extremely, discomfited, and/or insulted and/ or hurt and/or frightened	O'Driscoll (2020, p. 16)
"We label a post as offensive if it contains any form of non-acceptable language (profanity) or a targeted offense, which can be veiled or direct. This category includes insults, threats, and posts containing profane language or swear words."	Zampieri et al.,(2019)
"Obscene utterances, unlike other offensive uses of language, shock the listener entirely because of the particular words they employ, quite apart from any other message they may be intended to convey."	Feinberg (1983)

[8] Perspective API. (2024). About the API - Attributes and Languages. https://developers.perspectiveapi.com/s/about-the-api-attributes-and-languages

Systematically organizing and connecting how offensive language relates to other phenomena (e.g., misogyny, dangerous speech, homophobia, among others) is not an easy task – and again, there is no consensus on the matter. The interested readers are invited to check Poletto et al., (2021) for an initial attempt.

In the identification of hate speech and discriminatory discourse, the producer of the discourse and the individuals or groups targeted are among the most fundamental factors. The context in which the discourse is produced is also important in evaluating hate speech and discriminatory discourse. Although the basic characteristics of the groups that are the targets of hate speech are stated in a similar way in almost every definition, it is important to remember that these groups are not fixed and unchangeable, and that new groups may be targeted depending on context, time, and many other variables. It has been underlined that even though sex and/or gender is mentioned in many definitions as a target group characteristic, according to a survey released by the Council of Europe Youth Department, women remain a less visible target group (Council of Europe, 2016). The survey highlights that women are frequently targets of online hate speech, and this is often viewed as less severe than racial or religious hate speech. Participants in the survey acknowledged that sexist hate speech is a pervasive form of gender-based violence rooted in structural inequalities, reinforcing the broader patterns of discrimination faced by women. Furthermore, the study emphasized that framing gender equality and freedom of expression as conflicting values creates barriers to inclusive public discourse. In reality, gender equality and freedom of expression are mutually reinforcing, as true freedom of expression cannot exist when certain groups are systematically marginalized by hate speech.

When it comes to intersectionality and targeted groups, it would be no exaggeration to say that LGBTI+ people and refugees are among the groups most intensely exposed to hate speech in Turkey in recent years, in both print and social media. It is also possible to see that this situation is a reflection of the general political atmosphere, especially the discourse produced by political parties during election periods. Based on this, Eser Selen makes a direct connection between hate speech and violence as she analyzes how LGBTI+ refugees in Turkey are exposed to hate speech and how they are affected by it.

Intersectionality and hate speech: the plight of LGBTI+ citizens and migrants in Turkey

Eser Selen

The concept of intersectionality (Crenshaw, 1989; Acker, 2012) is crucial for understanding how social categorizations – such as race, ethnicity, religion, language and class – intersect with sexual orientation and gender identity, resulting in compounded experiences of hate speech (Chun et al., 2013) that incites active violence as a form of discriminatory discourse. Hate speech, by encompassing xenophobia, racism and other forms of prejudice and stigmatization, perpetuates marginalization and violence against LGBTI+[9] individuals, hindering their integration and full participation in society (Faloppa et al., 2023). In Turkey, LGBTI+ citizens and migrants navigate a complex landscape of intersecting vulnerabilities. Hate speech compounds the challenges they face based not only on their sexual orientation and gender identity, but also on factors like race, ethnicity, religion, language, class and national origin or migrant status. For instance, a gay Syrian refugee in Turkey may face prejudice based on their nationality, language, religion and sexual orientation, amplifying their vulnerability to hate speech and other forms of violence.

Human rights advocates, policymakers and researchers need to consider the intersectionality of factors including but not limited to ethnicity, race, religion, gender and sexual orientation when addressing hate speech, as these factors can compound the harm caused by such speech. Recognizing how hate speech impacts individuals with intersecting identities is essential to understanding its full scope. Furthermore, categorizing hate speech is crucial for a practical response and to counter hate speech in online and offline contexts. This is particularly vital for LGBTI+ migrants who face compounded challenges concerning their subjectivities. Different groups experience hate speech in unique ways, and detailed categorization would allow for tailored approaches to address specific forms (Yuying et al., 2013). However, hate speech is not always overt and vocal; it can manifest through various means, including memes and subtler forms of expression, i.e., microaggressions (Nadal, 2008; Nadal et al., 2011a; 2011b).

[9] In Turkey, activists and those in the community use the abbreviation LGBTI+ to identify themselves so I have also adopted that abbreviation for this report.

It is also important to reiterate that the definition of "hate speech" can be elusive and vary across contexts (Russell, 2020). However, even if the content of the discriminatory discourse does not meet any "accepted" definitions, individuals still have the right to perceive it as hate speech, and not just hateful communication, based on the discourse's harmful and offensive nature. The perceptions of LGBTI+ individuals' and their responses to societal discourses and actions are deeply influenced by the pervasive phobia surrounding their gender identity and sexual orientation. This phobia acts as a critical factor in the discrimination matrix, shaping how they navigate societal attitudes, behaviors and emotional responses (Wijaya, 2022). Furthermore, the internalization of phobia concerning gender identity and sexual orientation varies within the LGBTI+ community, impacting self-perception and well-being (Gao et al., 2023; McLean, 2021). Recognizing the impact of internalized phobia on whether individuals perceive such discourse as hate speech is essential for fostering inclusivity. Additionally, it is essential to acknowledge that lesbian, gay, bisexual, trans,[10] questioning, queer, intersex, and non-binary experiences cannot be evaluated in a vacuum (Fassinger & Arseneau, 2007). Therefore, close attention must be paid to the full spectrum of LGBTI+ individuals' identities and experiences.

LGBTI+ individuals face multifaceted challenges and barriers in accessing basic social and economic rights worldwide (Alessi et al., 2017; Williamson, 2023). Research showcases the myriad ways that LGBTI+ individuals – migrant or otherwise – have experienced stigma in terms of employment, housing, healthcare, and civil and family law, and also how they have been victims of hate speech, AIDS-related discrimination, police harassment, and violent hate crimes (Aswad & Kaye, 2020; Edward et al., 2020; Gowin et al., 2017; Hopkinson, 2017; Morales, 2013; Piwowarczyk, 2017). Prejudices, stereotypes, misinformation, and disinformation further contribute to the challenges faced by LGBTI+ citizens and migrants. Among these issues, violence against LGBTI+ individuals undoubtedly takes precedence since heterosexualism, as the moral (sexual) compass of society, resorts to public morals as an excuse – or justification – for acts of violence, which are often fatal (Selen, 2020). In Turkey, for example, "normal/ized" violence against women has regularly been condemned in the mainstream media by state officials, influential journalists, and celebrities who almost always exclude violence against trans women (or any LGBTI+ citizens) from their

[10] My use of the term "trans" is an inclusive identifier, rather than an identification encompassing individuals who identify as transgender, non-binary, agender, gender-fluid, and others, affirming the diversity of gender experiences.

discourses (Selen, 2020). The normalization of violence against LGBTI+ individuals is fueled by government officials' gender ideology and anti-LGBTI+ rhetoric including hate speech has created a dangerous environment impacting every facet of their lives (Selen, 2020)[11]. Additionally, the potential infiltration of gender ideology discourse into the legal system through discriminatory changes to the constitution and laws poses a severe threat (McClain & Waite-Wright, 2015). This deeply ingrained issue is further exacerbated by government policies condoning such acts (McClain & Waite-Wright, 2015).

In a similar vein, hate speech against migrants and refugees is a pressing global issue, and Turkey is no exception (Badali, 2019; Eduardo et al., 2013; Filibeli et al., 2021; Özdüzen & Korkut, 2020). Hate speech directed towards migrants is a central and growing theme in the context of Turkey. The current public narrative often portrays migrants as a "problem" requiring a solution. According to the United Nations High Commissioner for Refugees (UNHCR), Turkey maintains its position as the country with the highest number of refugees worldwide, hosting a substantial number of individuals displaced due to conflict, violence and persecution on a global scale. Presently, Turkey accommodates approximately 3.6 million Syrian refugees and around 320,000 individuals from other nationalities who are considered persons of concern (UNHCR, 2024; Kaya, 2023). Research demonstrates the severe consequences of discriminatory language and behavior, including hate speech, directed towards LGBTI+ asylum seekers and migrants worldwide (Alessi et al., 2017, Aswad & Kaye, 2020; Edward et al., 2020; Gowin et al., 2017; Hopkinson, 2017; Morales, 2013; Piwowarczyk, 2017). In the context of Turkey, where LGBTI+ rights have encountered significant obstacles in recent years, hate speech against LGBTI+ migrants can worsen existing discrimination and hostility towards the community, affecting their wellness, health, livelihood and safety.

The rising wave of anti-LGBTI+ hate speech in the digital landscape exposes the prevalence of hate speech and reveals the following:

[11] It is safe to say that since the Gezi Protests in 2013, LGBTI+ people have become the one of the most targeted minorities in Turkey. LGBTI+ individuals are targeted not only by the anti-LGBTI+ movement in the country but also by high-ranking government officials and political party leaders as a part of their gender policies, as well as influential journalists who have been in play for over two decades (see Selen, 2020). Additionally, violations of the right to peaceful assembly severely limit LGBTI+ visibility in the public sphere, further exacerbated by censorship across broadcasting, social media, and the arts (see Selen, 2012; Selen, 2020; Kılıç, 2023).

- **Lack of Accountability:** The alarming absence of official investigations into violent attacks and murders targeting lesbian, gay, bisexual, and trans individuals demonstrates a systemic failure to protect and deliver justice.
- **Forced Invisibility:** Pressure to remain closeted creates a climate of fear, where LGBTI+ individuals constantly risk their safety and/or lose their livelihoods without any legal recourse.
- **Dehumanizing Rhetoric:** The pervasive association of LGBTI+ subjectivity with "perversion", "disease", "terror" and "threat" serves to deny the existence and rights of LGBTI+ citizens and migrants.

By examining the online narratives and connecting them to real-world experiences, we can expose the interconnected nature of hate speech and its devastating impact on the lives of LGBTI+ individuals in Turkey (Özdüzen & Korkut, 2020). This analysis is necessary for developing effective strategies to combat hate speech and create a more inclusive society.

Data sets created through text mining from social media sites such as Twitter (2002-2020) or Facebook (2019-2020) revealed a preliminary Boolean search for terms like "LGBT*", "trans*", "homosexual", and "bisexual" in relation to terms such as "immoral", "pervert", "perversion", "disease", "threat", and "terrorist." These terms indicated not only hate and/or dangerous speech such as "eradication", "killing", "annihilation", and "damnation", but also other forms of "discriminatory discourse" against LGBTI+ individuals in Turkish (Table 3).

In addition to safeguarding and actively supporting the essential work of LGBTI+ organizations, researchers should communicate accessible analyses to the public regarding LGBTI+ rights and issues to achieve equality and ensure the protection of social and economic rights. The following measures are crucial in eliminating hate speech:

- **Counter hate speech:** Actively counter hate speech against LGBTI+ citizens and migrants in the digital landscape by raising awareness and promoting inclusivity.
- **Provide support services:** Offer specialized support services and resources tailored to the needs of LGBTI+ citizens and migrants, such as mental health counselling, legal aid, and assistance with housing and employment.

- **Empower communities:** Empower LGBTI+ communities and advocate for their rights through education, capacity building and legal aid.
- **Strengthen legal protections:** Expand and strengthen legal protections against hate speech and discrimination to explicitly include LGBTI+ individuals.
- **Ensure accountability:** Hold hate speech perpetrators accountable through legal action and public condemnation.
- **Enforce legal protections:** Ensure effective enforcement and implementation of legal protections against hate speech.

Table 3. Examples of database entries

User Handle	Keyword Search	Platform	Post	Age	Gender
@anonymized_#1	LGBT*	Twitter (2020)	"LGBT= perversion, the corrupted state of humans; they should be exterminated, they should be executed. Where there is disease, germs are killed; the LGBT community is the germ that makes people sick."	Unspecified	N/A
@ anonymized_#2	Trans*	Facebook (2019)	"May God give them trouble. May my Lord devastate, *al-khair* [damn] them. Let them be destroyed as soon as possible."	28	M

The Hrant Dink Foundation's decade-long monitoring of hate speech in Turkish print media (2009-2019) has revealed that while the ranking of targeted groups fluctuates with shifting political and social dynamics, certain identities remain persistent targets of hate speech regardless of current events. **Armenians, Jews, Christians, Greeks, and Greek Cypriots** are found to have been constant targets of hate speech in print media, with hostility toward these groups rooted in deep-seated historical narratives and reinforced through political discourse. HDF refers to these groups as **"the unchanging others"**, emphasizing how their portrayal in media remains consistent over time. Following the mass migration from Syria in 2014, **Syrians** rapidly became one of the most frequently targeted groups. This shift underscores how hate speech discourse continually constructs and reinforces **"new others"**, expanding beyond historical targets to include groups framed as contemporary threats.

Analysis of the types of hate speech directed at these groups highlights recurring patterns. Across all six groups, the most prevalent forms of hate speech have been *exaggeration, attribution and distortion*, where misinformation and negative generalizations perpetuate stereotypes and existing biases. Each group, however, is targeted through distinct narratives. Armenians are disproportionately subjected to *enmity and war discourse*, in which they are portrayed as "internal or external threats to national identity". Greeks and Greek Cypriots are often placed within interwoven narratives of "hostility", shaped by long-standing media portrayals that depict them as "adversaries". Syrians, in contrast, are primarily targeted through a rhetoric of "demographic and cultural threat", with media discourse amplifying fears of "erosion of national identity, crime and political instability" under the guise of economic concerns.

In recent years, monitoring efforts have expanded to include hate speech against **LGBTI+**s, revealing that sexual orientation and gender identity have become central axes of discriminatory discourse. Although the original study focused on ethnic, national and religious identities, LGBTI+s have increasingly been constructed as a "new other", targeted through state policies, official rhetoric and a hostile political climate. Far-right, pro-government mainstream media outlets play a key role in reinforcing this discourse, portraying LGBTI+s as “deviants” and framing their existence as a “societal threat”.

HDF’s monitoring work continues to document and analyze these patterns, demonstrating how hateful narratives and discriminatory rhetoric remain deeply embedded in Turkey’s media landscape. By tracking the evolution and consistency of hate speech, these findings provide critical insights into structural discrimination and the role of the media in shaping public perceptions of targeted groups.

HATE SPEECH DETECTION: POLICIES, METHODS AND CHALLENGES

Efforts to detect hate speech exist within a broader landscape shaped by legal frameworks, platform policies, and an evolving societal understanding of harmful discourse. Policies and frameworks addressing hate speech vary across legal and digital landscapes, shaped by cultural, political, and technological factors. While state-level policies often rely on legal mechanisms to regulate and penalize hate speech, social media platforms establish their own frameworks to moderate content within their ecosystems. These approaches differ in scope, enforcement, and underlying motivations, reflecting both the complexities of defining hate speech and the broader tensions between combating harmful discourse and preserving freedom of expression.

These policies influence how detection methods are developed from manual human annotation to AI-driven approaches. However, detection is also shaped by technological capabilities, linguistic nuances, and cultural contexts that are not always accounted for in legal or platform-level regulations. In some cases, policies may even hinder detection efforts—either by imposing restrictive definitions that fail to capture certain forms of harmful discourse or by creating broad, vague criteria that lead to over-censorship and suppression of speech.

This section begins by examining state and platform-level policies on hate speech to understand how they interact with detection methodologies. It then explores the different approaches to detecting hate speech, including human annotation, dataset curation, and AI-based models, while addressing the ongoing challenges of bias, accuracy, and ethical considerations. Recognizing the interplay between policies and detection methods allows for a more nuanced discussion on the complexities of identifying hate speech in practice.

Understanding policies on hate speech is essential for critically assessing their impact on detection efforts and the challenges they present. This section by Tirşe Erbaysal Filibeli examines the regulatory frameworks implemented by states and digital platforms, highlighting their enforcement mechanisms, limitations, and implications.

State policies, regulations and policy analysis

Tirşe Erbaysal Filibeli

Countries are adopting diverse approaches to regulate hate speech. The main problem while regulating hate speech is the blurry area between "hate speech" and "freedom of expression." According to the Media Pluralism Monitor 2024 Report (Bleyer-Simon et al., 2024), which evaluated 32 European countries, only Denmark, Finland, Germany, Lithuania and Sweden exhibit a low risk regarding protection against disinformation and hate speech. Conversely, Albania, Bulgaria, Cyprus, Hungary, Malta, Montenegro, Romania, Serbia, Slovenia and Turkey have underdeveloped and non-inclusive hate speech policies, creating a high risk to media pluralism and diversity.

For the most part, social media platforms have adapted U.S. laws and regulations in their terms of service. However, for the U.S., there is no legal definition of hate speech; most forms of hate speech are protected under the First Amendment of the Constitution to protect freedom of expression.

Germany and France have specific laws to fight online hate speech. Germany has a strict hate speech policy that criminalizes Holocaust denial, incitement to hatred and dissemination of Nazi propaganda. Germany's Network Enforcement Act was amended in 2020, and its Protection of Minors Act was amended in 2021 with the aim of combating online hate speech and protecting against online harm (Holznagel et al., 2023). In 2020, France adopted the Avia Law, which required online platforms by law to remove hate speech within 24 hours, after being noticed; this law, however, raised concerns about its possible negative effects on freedom of expression (Rebillard & Sklower, 2022). Although Denmark does have legislation against hate speech, there is no explicit mention of online hate speech (Simonsen, 2023). In Lithuania, there is a unique and collaborative system in which hate crimes and hate speech can be reported to online platforms created by public institutions and NGOs (Balčytienė & Juraitė, 2022).

In Turkey, on October 18, 2022, the government published Law No. 7418 on the Amendment of the Press Law and Certain Laws in the Official Gazette. The legislation amended Article 217 of the Turkish Penal Code, under the section on "Offences against Public Peace" by adding a clause that criminalized "publicly disseminating misleading information." However, in this law there is no clear definition of disinformation and hate speech, therefore leaving the understanding of these terms up to the interpretation of the legal authorities (İnceoğlu et al., 2022).[12]

The lack of a clear definition of hate speech can lead to the arbitrary enforcement of laws. This situation negatively impacts journalists, human rights activists, academics and other advocates. Therefore, it is essential to develop comprehensive and inclusive hate speech policies to prevent such arbitrary actions.

[12] **Editorial Note:** Building on the legal ambiguity introduced by Law No. 7418, the recent *Etki Ajanlığı* (Agent of Influence) draft law, proposed in January 2024, further expands the scope of vague legal definitions, raising concerns about its potential impact on multiple sectors. The draft law could criminalize individuals and organizations receiving foreign support, posing significant risks to freedom of speech and civil society. Critics argue that its ambiguous wording allows for arbitrary implementation, potentially targeting journalists, academics, and NGOs. If passed, the law could further restrict independent voices and international collaboration, reinforcing an atmosphere of self-censorship and legal uncertainty.

Social media platforms' current policies on hate speech

Tirşe Erbaysal Filibeli

There is an ongoing debate about the regulation of social media platforms, with discussions on their accountability for the content shared and the societal impact of their algorithms and policies. For instance, policies on hate speech generally focus on prohibiting content that promotes or encourages violence, discrimination, harassment or hatred based on characteristics such as race, ethnicity, religion, gender, sexual orientation, disability or other protected categories. But policies vary among social media platforms.

According to Statista,[13] the top five most popular social networks worldwide as of April 2024, were Facebook, YouTube, WhatsApp, Instagram and TikTok. In Turkey, the most used social media platforms were Twitter, Instagram, Facebook, WhatsApp and TikTok. As the most widely used platforms, these are also the forums most commonly used to spread hate speech and disinformation.

The hate speech policies of social media platforms have a common goal, which is to **create a safe environment for users balanced with freedom of expression**. However, the implementation and specifics of the platforms' policies vary significantly.

Meta: Facebook, Instagram, Threads, and WhatsApp

Meta's[14] hate speech policy for Facebook, Instagram and Threads aims to create a safe space by prohibiting attacks based on protected characteristics like race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity, and serious disease.

Meta's policy on hate speech is structured into two tiers, to balance free expression with protection against hate speech:

[13] **Statista.** (2024). *Most popular social networks worldwide as of April 2024, by number of monthly active users*. Statista. https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

[14] **Meta.** (2024). *Hate speech policy – Community standards*. Meta Transparency Center. https://transparency.fb.com/en-gb/policies/community-standards/hate-speech/

- Tier 1 defines severe violations as violent speech or advocacy or glorification of violence against individuals or groups based on protected characteristics, dehumanizing speech, including comparisons to animals, pathogens, or other sub-human entities, and promoting harmful stereotypes.
- Tier 2 defines statements of inferiority and contempt as claims about groups being physically, mentally, or morally deficient, including derogatory terms related to hygiene, appearance, intelligence, education, mental health, sexual behavior, expressions of hate, disgust, or dismissal towards protected characteristics, and targeted cursing aimed at insulting or degrading individuals based on these characteristics, except in certain contexts such as romantic breakups. It also defines exclusion and segregation, including content advocating for or supporting calls for action or statements advocating for the separation of groups based on protected characteristics or statements that explicitly call for exclusion, such as denying access to political participation, economic entitlements, or social spaces.

Meta determines specific types of prohibited content, including dehumanizing language and harmful stereotypes, and its policy on hate speech bans slurs attacking individuals based on protected characteristics, but it acknowledges exceptions for content shared to condemn hate speech, raise awareness, or when slurs are used self-referentially or in an empowering way. However, intent must be clearly indicated; ambiguous intent may lead to content removal. Furthermore, satirical content may be exempt if it criticizes or mocks the violating elements.

Instagram

In the Community Guideline of Instagram[15], there is a subtitle "Respect other members of the Instagram community." Here, they determine that *"to foster a positive diverse community, they remove credible threats or hate speech, content that targets private individuals to degrade or shame them, personal information meant to blackmail or harass someone, and repeated unwanted messages."*[16] For credible threats and hate speech definitions, Instagram gives a hyperlink to Meta's webpage.

[15] **Meta.** (2024). *Hate speech*. Facebook Help Center. https://www.facebook.com/help/instagram/477434105621119

[16] **Meta.** (2024). *Bullying and harassment policy – Community standards*. Meta Transparency Center. https://www.facebook.com/communitystandards/bullying

Instagram's policy determines that they allow stronger conversation, when advocating against hate speech through showcasing harmful examples, they may allow it by asking users to express their intent clearly.

Tiktok

TikTok defines its policy on hate speech under the title "Safety and Civility"[17] in its Community Guidelines (December 2024). The platform's policy underlines that being civil involves recognizing everyone's inherent dignity and being respectful in words, actions and tone. Within the section on Safety and Civility, there is a further subsection on "Hate Speech and Hateful Behaviors", which states that TikTok values the diverse backgrounds of its community members and prohibits hateful behavior, hate speech or the promotion of hateful ideologies. The policy encompasses content that attacks individuals or groups based on protected attributes such as caste, ethnicity, national origin, race, religion, tribe, immigration status, gender, gender identity, sex, sexual orientation, disability, or serious disease.

The policy also covers hateful ideologies and systems of beliefs – such as racial supremacy, misogyny, anti-LGBTQIA+ and antisemitism – that discriminate against individuals based on protected attributes.

TikTok defines prohibited content as in the following:

> Promoting violence, segregation, discrimination, and other harms based on a protected attribute; promoting any hateful ideology or claiming supremacy over a group of people based on protected attributes; demeaning someone based on their protected attributes; using hateful slurs associated with a protected attribute; denying historical events that harmed groups based on a protected attribute; blaming an entire protected group for the harmful actions of one individual who shares that attribute; sharing content that dehumanizes or invalidates people based on their protected attributes intentionally targeting transgender or gender non-conforming individuals by deadnaming or misgendering, facilitating the trade of items that promote hate speech or hateful ideologies.

As Meta, there are allowed content such as "*self-referential slurs*" and "*content that features statements intended to criticize or report on hateful speech*".

[17] **TikTok.** (2024). Safety and civility guidelines. TikTok Community Guidelines. https://www.tiktok.com/community-guidelines/en/safety-civility/

In its Community Principles, TikTok also includes a subsection on "Violent and Hateful Organizations and Individuals", which bans from the platform *"the presence of violent and hateful organizations or individuals (…) includ[ing] violent extremists, criminal organizations, violent political organizations, hateful organizations, and individuals who cause serial or mass violence."*

X (Formerly Twitter)

Within the "Rules and policies" of X (formerly Twitter), hate crime is covered within the section on Safety and Cybercrime, under the title "Hateful Conduct".[18]

As with the previous policies listed above, X's policy states that it is designed to protect individuals and groups from attacks based on race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability or serious disease. However, it differs from other policies by making specific mention that the platform aims to foster free expression and a public conversation inclusive of diverse perspectives. The policy sees hate speech as an abuse of rights and provides a link to report abuse[19] if users believe someone has violated X's policy.

X states that the platform will *"review and take action against reports of accounts targeting an individual or group of people with any of the following behaviour [i.e. behaviour in violation of its policy], whether within Posts or Direct Messages."* The policy then goes on to provide details on policy violations with definitions of hateful references, incitement, slurs and tropes, dehumanization, hateful imagery and hateful profile:

- **Hateful references** target individuals or groups with content that references violence or violent events where a protected category was the primary target, with the intent to harass. This includes references to genocides and lynchings.
- **Incitement** refers to behavior that incites harassment or discrimination, on or off the platform, against protected categories, including spreading fear or stereotypes. The definition also gives a direct link to the platform's "Violent Speech Policy."[20]

[18] **X (formerly Twitter)**. (April 2023). Hateful Conduct. X Help Center. https://help.x.com/en/rules-and-policies/hateful-conduct-policy

[19] **X (formerly Twitter)**. (April 2023). *Report abuse – Safety and sensitive content policies*. X Help Center. https://help.twitter.com/en/forms/safety-and-sensitive-content/abuse

[20] **X (formerly Twitter)**. (2024). *Violent speech policy*. X Help Center. https://help.twitter.com/en/rules-and-policies/violent-speech

- **Slurs and tropes** include content that degrades or reinforces negative stereotypes.
- **Dehumanization** refers to dehumanizing portrayals based on religion, caste, age, disability, serious disease, national origin, race, ethnicity, gender, gender identity or sexual orientation.
- **Hateful imagery** refers to logos, symbols or images promoting hostility based on protected characteristics, the use of which is banned on the platform. This includes historical hate symbols and images manipulated to include hateful symbols.
- **Hateful profile** refers to the use of hateful images or symbols in profile images, headers and user bios, including engaging in targeted harassment or expressing hate towards protected categories.

The platform's policy clarifies that actions taken against violations vary based on the severity and the individual's history of rule violations, ranging from making content less visible, to suspending accounts. The policy states that X requires reports of violations and offers an appeal process for suspended accounts.

YouTube

YouTube's policy on hate speech prohibits content that promotes violence or hatred against individuals or groups based on various protected attributes such as age, caste, disability, ethnicity, gender identity and expression, nationality, race, immigration status, religion, sex/gender, sexual orientation, victims of a major violent event and their kin, and veteran status.

YouTube gives a direction to report violations[21] and also guides users on how to report them. Unlike the guidelines of other platforms, and in line with its being a video-sharing website, YouTube has a video description of its hate speech policy.

YouTube's policy prohibits posting content that encourages violence or threatens individuals or groups due to their protected status (mentioned above), incites hatred against them, or dehumanizes them by comparing them to non-human entities. It also bans content that praises or glorifies violence against protected groups, uses slurs or stereotypes to promote hatred, asserts the superiority or

[21] **YouTube.** (2024). *Harassment & cyberbullying policy*. YouTube Help. https://support.google.com/youtube/answer/2802027

inferiority of any group, promotes hateful supremacism or recruitment for such ideologies, spreads conspiratorial claims targeting certain groups, denies or trivializes significant violent events or their victims, or attacks people based on their emotional, romantic, or sexual attractions. This policy extends to videos, video descriptions, comments, live streams, and any other product or feature offered by YouTube, including external links in content.

Declaring itself as a platform for free expression, YouTube makes "exceptions for videos that have a clear educational, documentary, scientific or artistic purpose"[22]:

> This would include, for example, a documentary about a hate group; while the documentary may contain hate speech, we may allow it if the documentary intent is evident in the content, the content does not promote hate speech, and viewers are provided sufficient context to understand what is being documented and why.

Violations of this policy will result in content removal, and the content creator will be notified by email. First-time violators may receive a warning without penalty, however they have to take policy training to allow the warning to expire after 90 days.[23] If violations continue, the user is subject to more severe actions,

Table 4. Summary of social media platforms' policies on hate speech

Platform	Policy Highlights	Content Prohibited	Allowed Content	Enforcement	Additional Notes
Facebook	Prohibits dehumanizing language, harmful stereotypes; structured into three tiers.	Tier 1: Violent, dehumanizing speech; Tier 2: Statements of inferiority.	Exceptions for awareness, condemnation of hate speech, or self-referential use.	Tiered enforcement based on severity; potential content removal.	Specifics provided for different types of violations.

[22] **YouTube.** (2024). *Standing up to hate: Our policies and exceptions*. YouTube. https://www.youtube.com/intl/en_us/howyoutubeworks/our-commitments/standing-up-to-hate/#policy-exceptions

[23] **YouTube.** (2024). *Hate speech policy*. YouTube Help. https://support.google.com/youtube/answer/2801939?ref_topic=9282436&hl=en-GB&sjid=1470727164900891766-EU#zippy=%2Cmore-examples

Instagram	Prohibits attacks based on protected characteristics; promotes a positive, diverse community.	Credible threats, hate speech targeting private individuals, sharing personal info to harass.	Allows discussions raising awareness about hate speech with clear intent.	Removal of violating content; user reporting system.	Links to Meta's general policy for definitions.
WhatsApp	Limited monitoring capability due to encryption; focuses on user reporting.	Any content spreading hate speech as reported by users or detected via suspicious activity.	Not applicable, as direct content monitoring is not possible.	Bans accounts based on user reports and detection of terms of service violations.	Encrypted communication limits direct content moderation.
TikTok	Prohibits hateful behavior, speech, and promotion of hateful ideologies.	Content attacking individuals or groups based on protected attributes, hateful ideologies, denying historical events.	Self-referential slurs, educational content on hate speech.	Bans hateful organizations and individuals, violent extremists, and perpetrators of mass violence.	Includes additional guidelines depending on the specific issue (e.g. human trafficking, harassment and bullying.
X (Twitter)	Protects against attacks based on various protected characteristics; fosters free expression.	Incitement, hateful references, slurs, dehumanization, hateful imagery, and hateful profile information.	Not explicitly detailed, focuses more on what is prohibited.	Actions range from making content less visible to suspending accounts.	Includes a process for appeals and reports of violations.

YouTube	Prohibits content promoting violence or hatred based on protected characteristics.	Incitement to hatred, violence, dehumanizing portrayals, hateful supremacism, denial of significant events.	Educational, documentary, scientific, or artistic content with clear context.	Content removal, warning system, possible channel termination and imposing monetization restrictions.	Extends to all features including comments and live streams.

including channel termination and suspension of access to ads and other monetization features. YouTube may also limit some features such as comments, suggested videos and likes for content that does not violate its policies, but is close to the removal line or could be offensive to some viewers.

The policies referenced in this report reflect the most recent policies of these platforms at the time of writing; however, during the preparation of this publication, Meta updated its policies three times, and YouTube and TikTok also implemented significant changes. In early January 2025, Meta CEO Mark Zuckerberg announced significant policy changes regarding hate speech and third-party fact-checking on the company's platforms that have sparked widespread debate. Zuckerberg stated that Meta would discontinue its third-party fact-checking program, transitioning to a "Community Notes" system, similar to the approach used by X. His speech raised concerns among experts and civil society organizations about the platform's ability to mitigate the flow of disinformation and hate speech. The proposed "Community Notes" could lack the accountability of professional fact-checking, leaving the platforms more vulnerable to campaigns of hate and disinformation.

The backdrop of these developments includes a politically charged atmosphere in the United States, following the re-election of Donald Trump as president. Images of prominent tech CEOs, including Zuckerberg and Elon Musk, standing behind Donald Trump at his inauguration have further fueled apprehension about the platforms' alignment with government priorities. The images make anyone who has concerns about the policies of those platforms wonder who will ultimately rule the platforms, and raise serious questions about the degree of independence these companies maintain in crafting policies and enforcing measures related to hate speech and disinformation.

Having examined different policies on hate speech, we now turn to the practical challenges of identifying and analyzing it. Addressing hate speech effectively requires not only an understanding of the policy measures but also robust detection methods that account for subjectivity, accessibility, and contextual nuance.

Considering the variety of target groups of hate speech and its dependency on context, one of the most important debates is how to detect and classify hate speech. Even when working with data sets that adhere to comprehensive international definitions, determining whether a given discourse constitutes hate speech can be difficult. This challenge is amplified by the need to account for numerous factors, including the context in which the discourse is produced, the identity of the individual responsible for its creation, the power dynamics it reflects, the audience it targets, the medium of dissemination, the tone of delivery and how it is perceived, along with its potential consequences. Additionally, elements that are harder to quantify, such as the cultural codes of the society in which the speech is produced and the historical connotations associated with certain words, play a significant role in detecting hate speech.

Researchers employ different categorizations to analyze hate speech, yet the detection process still relies on human labor. Even manual analysis by those who understand the cultural context and power dynamics can yield subjective results. Additionally, as social media platforms are increasingly used, they have also become places where hate speech is frequently seen. Considering the volume of content shared on these platforms, manual analysis is increasingly becoming insufficient (or even impossible) for working in these areas. It is this that has led to the need (or obligation) to resort to technology. However, automating the detection process through artificial intelligence (AI) and machine learning also comes with its own challenges, both existing and new. For instance, the portability of models trained on a specific dataset can result in lower performance when applied to other datasets, even when the same definitions are adopted. These disparities arise due to differences in data distribution across dimensions such as time, platform and topic. As automation alone cannot fully grasp the nuanced and context-specific nature of such discourse, researchers using tools of natural language processing (NLP) and AI are always trying to find more accurate ways of detection, despite the rapid advancements in these technologies.

Progress in automating hate speech detection has also been hindered by limited access to high-quality datasets. These datasets are essential for training models, directly affecting their accuracy. However, acquiring such data is increasingly difficult due to restrictive social media policies and limited research resources. Even when datasets are well-annotated and yield high accuracy, their effectiveness can vary based on the context in which they are applied. A group targeted in one context may not be recognized as such in another, making it difficult to create models that are both accurate and broadly applicable. Moreover, bias in the models is another pressing concern. Trained models often mirror the data they have been trained on, which means that any annotation bias – such as the use of dialects or reappropriated slurs – can be amplified, further harming communities already targeted by hate speech. Additionally, many studies tend to focus on data relevant to dominant cultural or social groups, resulting in the underrepresentation of marginalized communities. This lack of diversity in data limits the ability of detection models to fully capture the scope of hate speech.

Explainability is yet another challenge. Approaches based on neural networks often lack transparency in their decisions, making it difficult to understand why a particular message is flagged as hateful. While solutions like feature-attributions or attention-based mechanisms have been introduced, their implementation remains problematic. Enhancing the transparency of model decisions is critical for both accuracy and ethical deployment, requiring further attention and research.

To address these issues, the following sections of this document aim to outline the usage of digital technologies in the detection of hate speech and other related concepts.

The impact of definitional variations of hate speech on datasets

Arzucan Özgür

The increase in hate speech on online social platforms makes it necessary to develop automated methods for recognizing and combating hate speech, since the amount of the data makes it almost impossible to work with manual methods. The quality and reliability of automated methods is highly dependent on the datasets used to train these models. Annotating datasets to detect hate speech is a difficult task due to the potentially subjective nature of the concept.

The difficulty of defining hate speech also makes it difficult to write annotation guidelines. Unclear guidelines can further increase the subjectivity of annotations and lead to biased datasets and models. Furthermore, it is difficult to achieve satisfactory inter-annotator agreement (Madukwe et al., 2020), leading to inconsistently labeled datasets. It has even been shown that the same tweets in the same dataset can be given different labels (Awal et al., 2020). One of the strategies used by researchers to mitigate this challenge is to include the data items (e.g., tweets) that multiple annotators agree on their labels. For example, in recent shared tasks — collaborative evaluations where teams work on the same problem — for hate speech detection in social media text, only tweets with at least three annotations are included in the training and test sets (Arın et al., 2023; Uludoğan et al., 2024).

Although there are a number of datasets annotated for hate speech, due to the different definitions of hate speech and annotation guidelines, models based on machine learning (ML) models that have been trained on one dataset have been shown to perform worse when tested on other datasets, demonstrating that different definitions of hate speech and different annotation guidelines limit the generalizability of the developed models (Swamy et al., 2019).

Human annotation of hate speech: Benefits and challenges

Ayşecan Terzioğlu, Didar Akar

Annotating hate speech is a nuanced and intricate process, influenced by multiple factors such as the limited availability of context of the message, variations in the annotation schemes, diversity of data aggregation methods and concerns regarding data quality. In the annotating process, disagreements about whether a statement contains hate speech or the degree of offensiveness in that statement are often attributed to the ambiguity of that statement or the annotators' personal opinions, biases and lived experiences (Novak et al., 2022).

These challenges have prompted scholars to develop performance measures that assess both the extent of annotator agreement and the overall quality of the detection model. Similarly, there has been a recent growth of academic literature focusing on the relationship between annotators' socio-economic backgrounds and annotation patterns. To mitigate the impact of these background differences, multiple sessions of training with a wide range of concrete examples are provided for the annotators. Annotators working in pairs or regularly coming together to discuss contentious statements are also other possible solutions to the problem (Kocon et al., 2021).

Although the absence of disagreement among annotators is often idealized as the "gold standard" in the annotation process, we know that it is impossible to reach that absolute consensus due to the annotators' diverse backgrounds and perspectives. Further complicating the inter-rater and even intra-rater reliability are factors such as the lack of adequate contextual information and the evolving nature of annotators' attitudes and opinions over time. Therefore, in order to minimize disagreements and enhance annotation consistency, clear and detailed guidelines accompanied by concrete examples are needed (Aroyo & Welty, 2015).

In projects aiming to develop AI-based solutions to hate speech, annotators' practices have a direct and substantial impact on the performance and accuracy of the resulting algorithms as it is the annotators' decisions that will form the training data set for those algorithms. However, the annotation process is neither simple nor straightforward. One of the most significant complicating factors is the interplay between the annotators' identity and their decision-making process.

The identity of the annotator is unsurprisingly a dimension that should be factored into the process. It is not a neutral element; it shapes how they perceive and categorize hate speech. An annotator's ethnic, political or familial background may make them more perceptive to certain types of hate speech while overlooking others. Similarly, their professional or educational backgrounds may lead them to consider different concerns in their decision-making. For example, an annotator who is familiar with the AI training procedure would consider the effects of annotating a not-so-clear case of hate speech fearing that it would train the algorithm in such a way that it would categorize neutral speech as hate speech, causing false positives. Thus, although objectivity as an epistemic stance is deemed desirable for annotators, it also comes with its disadvantages such as missing out on vaguely expressed hate speech in the absence of explicit or adequate contextual information.

Beyond differences in the annotators' backgrounds, another major obstacle to achieving annotator agreement is the differences in interpreting non-linguistic signs. As human beings, when we interpret a text, we use much more than the linguistic content. Other modalities such as visual elements in the form of images, memes, symbols and punctuation, emojis and so forth play a significant role in the meaning-making process. Moreover, world knowledge, intertextuality and implicit references are also put into use to understand the multiple levels of indexicalities the text activates.

These challenges highlight the need for an interdisciplinary approach to annotation, incorporating insights from linguistics, social sciences and computational fields to refine annotation methodologies and improve model performance.

Significance of annotator’s identity on hate speech detection models

Claudia von Vacano

Building a robust and transparent machine learning model involves tracking several key aspects, such as data representation, annotation, bias and explainability. Each aspect plays a critical role in ensuring the model’s reliability and fairness. The foundation of any good model is the data. It is crucial that the data used for training a machine learning model is representative of the real-world scenarios in which the model will operate. This means it should accurately reflect the diversity of cases and variables it is expected to encounter. Failure to achieve a representative dataset can lead to the amplification of biases. To address this, data collection must be approached with inclusivity in mind, including a wide array of data points from diverse demographics, conditions and variations. Techniques such as oversampling minoritized and marginalized groups can help achieve more balance.

Once a dataset has been curated appropriately according to the intended design, the next challenge is annotation. Annotators label the data that machine learning models learn from, and the annotators’ interpretations can introduce biases. These biases might stem from an annotator’s cultural background, race/ethnicity, religion, national origin/citizenship/immigration status, gender identity, sexuality, age, disability status or other protected categories for vulnerable populations. These unique identities inform a person’s personal experiences and shape their worldviews. In turn, these views influence how they perceive, interpret and label data, such as a given comment on a social media platform. *One annotator* may not understand a negative connotation based upon their personal and cultural context. This variability can affect the model’s learning process.

To address this, a powerful methodology is to combine a framework like Item Response Theory (IRT) with Deep Learning. Mark Wilson’s “Constructing Measures” approach provides a structured way to analyze the data produced by annotators. This approach helps in understanding the likelihood of an annotator labeling a particular item in a certain way based on the item’s characteristics and the annotator’s biases. By leveraging IRT, we can better understand and control for certain biases in contrast to other perspectives in the data-labeling process.

For example, if we have mostly white men labeling a dataset, we can strengthen the interpretations of minoritized groups in contrast. This can create an improved level of model sensitivity.

In "Assessing Annotator Identity Sensitivity via Item Response Theory: A Case Study in a Hate Speech Corpus" (Sachdeva et al., 2022), we explored the impact of annotator identity labeling patterns in a dataset used for training machine learning algorithms, particularly focused on hate speech identification and measurement. We used IRT to model and quantify how annotators' demographic identities influence their sensitivity to different types of hate speech.

We proposed a shift from viewing annotator judgments as biases, to understanding them as sensitivities. This reframing acknowledges the subjectivity involved in tasks like hate speech labeling, where annotators' personal identities can lead to different perceptions and interpretations of content.

IRT is used to quantify annotator sensitivity, providing a nuanced view that accounts for individual differences among annotators. This methodological choice enables us to assess the likelihood of an annotator assigning a specific label based on their identity, beyond a binary notion of right and wrong.

Our study examined over 50,000 social media comments labeled by approximately 10,000 annotators. Through the use of IRT, the study reveals that annotators tend to show increased sensitivity towards hate speech targeting groups they identify with. Three IRT techniques were used to measure sensitivity from different angles, revealing that annotators' race significantly correlated with their perceptions of hate speech targeting various racial groups. The analysis shows that African American annotators are more likely to identify comments as hate speech when these target Black individuals, compared to comments targeting other races.

The findings emphasize the importance of considering annotator identity in the design and implementation of machine learning models, as this can influence the training data and thus the behavior of algorithms. Understanding annotator sensitivity helps in mitigating algorithmic biases and fosters fairness. In future, researchers should explore how annotator identity interacts with other types of content and in other contexts. This study was based in the United States, but the same approach can be applied to different geographies. We advocate for better representation and consideration of diverse annotator identities in dataset creation and algorithmic design to enhance the fairness and accuracy of automated systems.

Over the past decade, the Hrant Dink Foundation has continuously refined its hate speech monitoring methodology, adapting to the growing complexity of media landscapes and the increasing role of digital platforms in shaping public discourse. Initially focused on print media, the Foundation's monitoring efforts provided a critical foundation for understanding the patterns and narratives of hate speech in Turkey. However, with digital spaces becoming the primary arenas for public discussion, the limitations of traditional monitoring methods is increasingly evident. The vast volume of online content, the speed at which harmful discourse spreads, and the evolving nature of hate speech – including coded and indirect expressions – necessitated the adoption of new methods and technological tools to detect and analyze hate speech more effectively.

The increasing volume and velocity of digital content made manual monitoring unsustainable, while the growing complexity of hate speech and its rapid spread demanded a more sophisticated and scalable approach. Recognizing this, the Foundation, in partnership with Boğaziçi University and Sabancı University, expanded their efforts to social media and AI-driven detection, developing a machine learning-based tool, ***pari***, to detect and analyze hate speech in Turkish and Arabic text.

This transition was not merely a technological shift but the result of extensive discussions, research, and refinement based on the Foundation's decade-long experience in hate speech monitoring. The categories of hate speech used in AI-driven detection serve as a structured framework for identifying hate speech while capturing its linguistic, social, and contextual nuances. The categories were developed through detailed discussions, interdisciplinary research, and a human rights-based approach, ensuring that they effectively capture the range of hate speech, while prioritizing the emphasis on freedom of speech.

The AI tool, ***pari***, developed as part of the *Utilizing Digital Technology for Social Cohesion, Positive Messaging and Peace* project, was trained through manual data annotation with a detailed hate speech monitoring methodology, consisting of four hate speech categories:

- **Exaggeration, Attribution, Distortion, and Generalization**

  Hate speech produced through negative generalizations, misinformation, or attributions that target an entire group based on the actions of individuals or isolated incidents.

- **Swearing, Insults, Degradation, and Dehumanization**

  Direct insulting, dehumanizing or degrading language used to target a group.

- **Enmity/War Discourse, Threat of Violence, Attack, or Harm**

  Discourse that portrays a group as a threat or enemy, fostering hostility and suggesting conflict or harm.

- **Symbolization**

  The use of the identity expression itself to negatively associate a group with an unwanted characteristic or to degrade their identity.

The shift to AI monitoring does not replace the need for human expertise. The Foundation continues to integrate human oversight to ensure accuracy, contextual understanding, and the ethical application of detection methods. Effectively detecting, monitoring, and analyzing hate speech requires a holistic approach that combines traditional research with digital innovation, allowing for more precise documentation and intervention while promoting inclusive, rights-based discourse in digital spaces.

Existing categories to detect hate speech via natural language processing

Roser Morante

Hate speech detection has become a popular Natural Language Processing (NLP) task (Fortuna & Nunes, 2018; Jahan & Oussalah, 2024). One of the first datasets was introduced in 2012 (Warner & Hirschberg, 2012), with more datasets being created in recent years (Poletto et al., 2021). Since the creation of those first datasets, a large number of new datasets have been created, likely a reflection of the fact that expressions of hate have become more prominent year upon year, reaching many sections of society.

How has hate speech been defined by the NLP community? In their paper "A Survey on Hate Speech Detection using Natural Language Processing", Schmidt and Wiegand (2017) choose the term "hate speech" because it acts as a broad umbrella term for numerous kinds of insulting user-created content. They adopt Nockleby's (2000) definition of hate speech as, "any communication that disparages a person or a group on the basis of some characteristic such as race, colour, ethnicity, gender, sexual orientation, nationality, religion, or other characteristic." Gibert et al., (2018), who present the first public dataset of hate speech annotated on internet forum posts in English at sentence-level, indicate that in most of the literature consulted (Nockleby, 2000; Djuric et al., 2015; Gitari et al., 2015; Nobata et al., 2016; Silva et al., 2016; Davidson et al., 2017) hate speech is defined as follows: a) a deliberate attack, b) directed towards a specific group of people, and c) motivated by actual or perceived aspects that form the group's identity.

Working on hate speech detection models, Röttger et al., (2021) opt for the following definition: "Abuse that is targeted at a protected group or at its members for being a part of that group." The protected groups covered by their research are different from the groups mentioned by other documents in this field: "We define protected groups based on age, disability, gender identity, familial status, pregnancy, race, national or ethnic origins, religion, sex or sexual orientation."

The European Union Agency for Fundamental Rights (2023) presents a typology of hate speech based on the hateful nature of the post content. The study establishes five main categories – not mutually exclusive – of online hate,:

- **Incitement to violence, discrimination or hatred:** refers to content that clearly encourages or urges the audience to commit violence; act in a discriminatory manner, which means treating someone differently because of a (perceived) protected characteristic; act in a hateful manner, including speaking or writing.
- **Denigration:** an attack on the capacity, character or reputation of one or more people in connection with their (perceived) membership of a particular group or, as the Council of Europe recommendation states, "by reason of their real or attributed personal [protected] characteristics".
- **Offensive language:** "hurtful, derogatory or obscene" language, such as insults referring to protected characteristics.
- **Negative stereotyping:** certain negative traits and characteristics are "negatively valenced and attributed to a social group and to its individual members" in relation to protected characteristics.
- **Other hateful content:** may include support for hateful ideologies or Holocaust denial.

Having defined hate speech, we should look at how it has been categorized. Within the NLP community, the approaches to annotating hate follow mostly three tendencies:

- Binary classification (hate vs. not hate).
- Hate speech vs. derogatory or offensive language.
- Different types of hate speech.

However, the examples that we will provide below will also show that depending on the research goals, the classifications can vary.

Warner and Hirschberg (2012) identified seven annotation categories to be assigned to paragraphs: anti-Semitic, anti-Black, anti-Asian, anti-woman, anti-Muslim, anti-immigrant or other-hate. They argue that forms of hate speech can be distinguished from each other by identifying the stereotypes being applied. According to the authors, each stereotype has certain language characteristics (one-word epithets, phrases, concepts, metaphors and juxtapositions) used to express hate. For example, anti-Hispanic speech might refer to border crossing or legal identification. Kwok and Wang (2013) follow this line by annotating tweets as racist anti-Black or not.

Waseem and Hovy (2016) classified a collection of approximately 16,000 tweets into racist, sexist or neither. Musto et al., (2016) collected geo-tagged data from Twitter to create a "Hate Map" showing the locations of five different types of hateful content: homophobic, racist, sexist, anti-Semitic and against disability, while Del Vigna et al., (2017) identified six categories: religion, disability, social status, politics, race, sex and gender issues and others. Gambäck and Sikdar (2017) assign tweets to one of four predefined categories: racism, sexism, both (racism and sexism), and non-hate-speech. The category "both" is important to show the intersectionality of the subject.

Gibert et al., (2018) present a public dataset of internet forum posts in English annotated at sentence-level as expressing hate or not. Sentences are annotated with the HATE label if they fulfill the three conditions already mentioned above: a deliberate attack, directed towards a specific group of people, and motivated by aspects of the group's identity.

Sharma et al., (2018) collected a set of 9,000 tweets containing harmful speech and they manually annotated them based on their degree of hateful intent, arguing that harmful speech exists on a spectrum of severity.

- **Class I:** Hate speech that is either public or directed at a particular group, mostly with no redeeming purpose. The authors argue that hatred and violent behavior projected at a group is stronger than individual accusations or violence. From the context, it is evident that the speaker intends to hurt sentiments of certain "isms" (extremism), potentially provoking a violent response in return.
- **Class II:** Cyber banter (accusing, threatening and using aggressive/ provocative language to disagree, etc.) and verbal dueling. Hate in this class is less intense than in Class I. It hurts sentiments, but not to the degree of invoking a violent response. Hate in this class can be highly provocative when addressing an individual rather than an ideology or community/group.
- **Class III:** Mildly provocative messages, mostly addressed to an individual entity, not necessarily targeting a group or community. This class uses more profane and filthy words, often in a context of trolling, irony, or sarcasm.

Salminen et al., (2018) manually label 5,143 hateful expressions posted on YouTube and Facebook videos and create a granular taxonomy with 13 main categories and 16 subcategories (29 in total). The main categories include four categories describing the type of language (accusations, promoting violence, humiliation, swearing) and nine describing the targets of the expressions:

- **Financial power:** hate toward wealthy people and companies and their privileges. Pointing out their intentions to manipulate and commit crimes.
- **Political issues:** hate toward government, political parties and movements, war, terrorism, the flaws of the system.
- **Racism and xenophobia:** racist comments toward black, white, Asian people. Generalizations about some characteristics, and hateful comments regarding refugees.
- **Religion:** everything about religion, including Judaism, Christianity, Islam, and religion in general, both as a subject or object of hatred.
- **Specific nation(s):** hate towards different countries, their systems, people (if the nationalities are mentioned), and certain events, like immigration, territory, and sovereignty.
- **Specific person:** hate toward specific people who can be regular people, politicians, millionaires, celebrities, or another individual related to specific news.
- **Media:** comments and emotional outbursts about bias and false statements made on purpose by the corrupted media.
- **Armed forces:** hate toward the military and law enforcement, and the way they operate, including unethical behavior.
- **Behavior:** hate toward the world, humanity, immoral actions of some part of the society, ignorant people, people that committed certain actions, and that have certain habits.

Bosco et al., (2018) describe a Twitter corpus of about 6,000 tweets, annotated for hate speech against immigrants. The annotation scheme includes hate speech, aggressiveness, offensiveness, irony, stereotypes, and (on an experimental basis) intensity. The authors considered two aspects for the identification of hate speech: 1) the target, which must be a group or an individual considered for its membership in that group (and not for its individual characteristics); 2) the action or illocutionary force of the utterance, such as spreading, inciting, promoting or justifying hatred or violence towards the given target, or aiming at dehumanizing, delegitimizing, hurting, or intimidating the target.

Fortuna et al., (2019) create a dataset for Portuguese composed of 5,668 tweets. First, tweets were assigned a binary label ('hate' vs. 'no-hate') by non-expert annotators. Expert annotators then classified the tweets following a fine-grained hierarchical multiple label scheme with 81 hate speech categories, including categories that are less commonly mentioned in hate speech classification, such as 'fat people', 'fat women', 'ugly people', 'ugly women', 'men', 'feminists', 'people with left-wing ideology'. Their work is important since they find that the inter-annotator agreement varies across the different categories, which indicates that some specific types of hate speech can be more difficult to classify than others.

Several datasets have been annotated for shared tasks. The Kaggle Toxic Comment Classification Challenge 2018[24] provided a dataset that contained 150,000 Wikipedia comments annotated for toxic behavior, within the following categories: toxic, severe toxic, obscene, threat, insult, and identity hate.

The HatEval (Basile et al., 2019) shared task, organized under the umbrella of the SemEval evaluation campaign, addressed the multilingual detection of hate speech against immigrants and women on Twitter. For this task, tweets were annotated as expressing hate towards women or immigrants or not, and the tweets that were classified as hateful were then further annotated as aggressive or not. For the second edition of the Hate Speech Detection task (Sanguinetti, 2020), which focused on detecting hateful content in Italian Twitter messages, tweets were annotated as either expressing hate or not, and also annotated with the stereotype being applied.

The series of EXISTS competitions (sEXism Identification in Social neTworks), held from 2021 to 2024 (Rodríguez-Sánchez et al., 2021, 2022; Plaza et al., 2023), focuses on detecting sexism in tweets. The 2024 edition[25] proposes several tasks: sexism identification (binary classification) and sexism categorization, as in previous editions. The sexism categories are ideological and inequality, stereotyping and dominance, objectification, sexual violence, and misogyny and non-sexual violence. Additionally, as in 2023, the task of source intention is included, which consists of categorizing the message based on the author's intention (direct, re-

[24] **Kaggle.** (2018). *Jigsaw toxic comment classification challenge: Overview*. Kaggle. https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/overview

[25] **UNED NLP Group.** (2024). *EXIST 2024: Explainable hate speech detection.* UNED. http://nlp.uned.es/exist2024/

ported or judgmental). Finally, in 2024, three new tasks have been introduced to address sexism in memes: sexism identification, sexism categorization, and source intention in memes.

The study of hate speech in NLP has evolved from binary classification to more complex, fine-grained categorizations. As hate speech continues to evolve, targeting new groups and emerging in different forms, refining annotation methodologies remains an ongoing challenge for researchers in the field.

Accessibility to data by researchers

Houda Bouamor, Onur Varol, Berrin Yanıkoğlu

The study of online interactions, content moderation, and the spread of harmful information, including hate speech, relies on researchers' access to social media data. However, this access is increasingly restricted due to platform policies, privacy concerns, and technical limitations. Social media companies regulate data availability through strict terms of service, proprietary algorithms, and evolving policies that often lack transparency.

Recent changes on platforms like X/Twitter, Facebook, Instagram, and Reddit have further limited researchers' ability to collect data. For example, X/Twitter restricts the number of messages that can be retrieved, while Facebook messages remain inaccessible to non-friends. These constraints, along with technological and regulatory barriers, significantly hinder researchers' ability to analyze online platforms effectively.

Data is difficult to access. Researchers find it difficult to obtain access to the necessary data to study hate speech on social media due to several factors. Social media platforms often protect user data closely, citing privacy concerns and ownership rights. Furthermore, the huge volume of user-generated content on these platforms makes manual data collection and analysis difficult and impractical, necessitating the use of automated approaches, which can be technically complex and resource-intensive. For instance, only a few platforms offer application programming interfaces (APIs), and some are not straightforward to use. Rate-limits for API requests and subscription-based models steer researchers towards collecting sample data or introducing strict criteria which eventually bias the datasets used in analysis. In 2022, a group of researchers challenged the rate-limits and API restrictions to capture the entire public stream of X/Twitter as a collective effort (Pfeffer et al., 2023). Despite such great effort, they were only able to capture a single day's worth of data, highlighting the challenges of collecting data at scale; under Elon Musk's ownership such practices became impossible as a result of API changes.

Access to data is regulated differently in different platforms. Researchers must comply with different regulatory requirements for each platform. Ethical considerations and regulations around data access and usage can vary significantly across different platforms. Some platforms can have stricter policies in place to protect user privacy, while others may have more flexible approaches. Researchers find themselves facing the challenge of interpreting these rules while applying them and dealing with the disparities. This can be a particular challenge if they want to work on different platforms for comparative research.

Data changes. Additionally, the dynamic nature of online content in these platforms, where data can be deleted by users or platforms, may erase or restrict access to certain data, making it difficult for researchers to study it in a consistent manner. Hate speech spreads very rapidly requiring real-time access to data by researchers, thus adding a layer of complexity. Similarly, social bot accounts weaponize these weaknesses, posting manipulative or hateful content and deleting them before the platforms or detection systems react; however, platform users have already consumed that information and are affected by them.

Users can game the platform rules and regulations. To protect users, platforms can detect hate speech automatically or give users capabilities to filter content based on predefined criteria. Accounts can also submit complaints about other users and their contents. However, especially in regards to hate speech, accounts can find creative approaches to get around these protective mechanisms, such as spelling certain words differently to get around filters or periodically posting and deleting content to reduce changes to acquire negative responses. For instance, the "right to be forgotten" is a provision within the General Data Protection Regulation (GDPR)[26] that allows individuals to request the removal of their personal data from online platforms under certain circumstances. Nevertheless, users might attempt to exploit "the right to be forgotten" as a shield for hate speech, attempting to erase their digital traces after engaging in harmful behavior online (Xue et.al, 2016). This poses a significant challenge for platforms and regulatory authorities trying to balance privacy rights with the need to combat hate speech and protect public safety.

The changing landscape of social media research due to data availability is concerning, especially at a time when we need research efforts towards detecting manipulation, coordinated activities, and systematic amplification of hate speech. Researchers are not equipped with the crucial data to quantify the problem, yet alone develop strategies to mitigate it.

[26] **GDPR.eu.** (2018). *General Data Protection Regulation* (GDPR). GDPR.eu. https://gdpr-info.eu/

The European Union's effort to introduce the **Digital Services Act**[27] (DSA) is a considerable effort to regulate social media platforms and introduces obligations on platform companies. Independent organizations such as the Center for Democracy and Technology[28] and the Coalition for Independent Technology Research[29] also support transparency and data access.

Bias. Beyond biases introduced during the data annotation process, as discussed in earlier sections, data collection from social media platforms also presents significant biases. While many platforms previously provided data access through APIs, the sampling strategies and potential biases inherent in these systems remain unclear. Studies on X/Twitter's API (Morstatter et al., 2013) have shown that publicly accessible random samples differ from those obtained via paid access, highlighting inconsistencies in data availability. Additionally, user behavior influences data collection. Since platforms impose rate limits on API access, developers optimize queries to maximize data capture, which in turn affects data quality: different sampling strategies can yield vastly different datasets. Concerns about data bias and platform transparency are well-documented. When asked about the datasets they would ideally use for misinformation research, most researchers emphasize these issues (Pasquetto et al., 2020).

[27] **European Commission.** (2022). *Digital Services Act*. European Commission. https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-act_en

[28] **Center for Democracy and Technology.** (2024). https://cdt.org/

[29] **Coalition for Independent Technology Research.** (2024). https://independenttechresearch.org/

Pending issues and potential solutions for developing tools for the automatic detection of hate speech

Tommaso Caselli

In the past ten years, Natural Language Processing (NLP) – the area of Artificial Intelligence that is focused on developing tools for the automatic understanding and generation of human languages – has seen a growing interest in the development of language resources (corpora and tools) for the automatic detection of hate speech and related phenomena. Previous work has been accompanied by shared tasks and competitions[30] covering many languages – although disparities are still present and only superficially tackled. On the basis of previous relevant surveys (Vidgen & Derczynski, 2020; Poletto et al., 2021), there are three key issues to discuss in relation to NLP: definitions, data sources, and data annotation.

Definitions

Since the challenges faced by researchers due to the lack of a universal definition of hate speech have already been discussed in earlier sections, we will not reiterate the same points but will address the topic briefly.

A potential starting point to steer the discussion towards a reference definition of hate speech is to acknowledge its subjective and offensive nature. Offensive language functions as a broad umbrella that can help carve out related phenomena such as hate speech. To this end a definition of offensive language proposed in two successful shared tasks (Zampieri et al., 2019, 2020) run at SemEval 2019 and 2020 – a series of workshops for the evaluations of NLP tools – can be of help. In these contexts, offensive language is defined as *any form of non-acceptable language, explicitly expressed or veiled, including targeted offenses, insults, threats and messages containing profane language* (see Table 1 under Offensive Language for the exact quote). The main advantage of this definition is its genericity, making it possible to apply it to many different instances. The fact that it targets *"any form of non-acceptable language"* is essential to address the subjective nature of these phenomena. The fact that the presence of a target is not compulsory offers a starting point to

[30] **Kaggle.** (2018). *Jigsaw toxic comment classification challenge*. Kaggle. https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge

distinguish hate speech from other phenomena. For instance, hate speech could be defined as a targeted insult towards an individual or a group because of inherent characteristics such as race, religion, sexual orientation, gender or political opinion, thus contributing to the identification of commonalities that could be further used in the development of annotation schemes and automatic tools. The possibility of reconnecting specific phenomena to their offensive nature is also a key aspect that can contribute to the development of a unified framework.

Data sources

Corpora for hate speech have been developed covering different language families, such as Indo-European languages, Turkic, Semitic, among others.[31] Nevertheless, it is impossible to identify a single commonly agreed reference dataset or benchmark in any language, making it difficult to compare tools and approaches that have been developed.

The variability in the sources of data is also a limitation. As highlighted by Vidgen and Derczynski (2020), Twitter (now X) has for the past 10 years been the major social media platform used to create datasets on this and related language phenomena. This was possible thanks to the openness of the platform's application programming interface (API) – until early 2023 – and its use by more than 300 million people in at least 34 supported languages. While this has had advantages in making it possible to create comparable datasets in many languages, it is also an intrinsic limit of the research conducted in this area. Focusing on one social media platform, which had specific characteristics concerning the maximum length of the messages[32] resulted in distorted, or biased representations of the language expressions that may trigger in a reader the offensiveness status of the messages. Besides the popularity of hate speech and related phenomena in NLP, the developed solutions are poorly portable as soon as the source of the data differs (e.g., a system trained on Twitter/X and applied on Reddit) even when the language phenomenon (e.g., data annotated for offensive language) does not change (Fortuna et al., 2020).

[31] **Zubiaga, L.** (2020). *The Hate Speech Dataset Catalogue: A comprehensive community-based resource documenting datasets for hate speech detection (and related phenomena) in multiple languages*. GitHub. https://github.com/leondz/hatespeechdata/tree/master

[32] **X.** (2024). *X Premium: Features and subscription tiers*. X Help Center. Retrieved from https://help.x.com/en/using-x/x-premium After the acquisition of Twitter by Elon Musk, the permitted length of posts varies according to the subscription plan: 280 characters for Basic tier users, and up to 25,000 characters for paid Premium and Premium+ tier users.

Social media clearly offers an ecological platform to collect natural instances of hate speech, but such platforms also constrain the release of the annotated data as well as the replicability of the experiments. For instance, the Twitter/X terms of use explicitly prohibit the public release of any piece of information except the ID of the message unless permission from the user – who is the sole owner of the content – is obtained. Although it is possible to retrieve the data by using techniques such as rehydration, i.e., re-downloading the messages that compose the dataset from the platform, the risk that some of the messages (and usually those particularly loaded with offensive content) have been removed by the platform or deleted by users is high, as mentioned in the previous section. This usually results in "shrinking" datasets: datasets whose size changes from one rehydration to the other, making it impossible to compare results across NLP systems. Solutions that attempt to mitigate, if not prevent, the impact of this problem implement anonymization of the messages by removing all user handles and URL links. This will avoid a user being able to identify the owner of a message if (and when) the message is removed from the social media platform, thus preserving the integrity of the collected data. This solution has been adopted for OffensEval 2020 (Zampieri et al., 2020), a shared task on multilingual offensive language detection held in the context of the SemEval workshop series. A similar solution has been adopted for the Dutch Abusive Language Corpus (DALC) (Ruitenbeek et al., 2022) which is anonymized and released upon the signature of a dedicated Data Sharing Agreement. Recently, an issue affecting the collection of data is the increasing restrictions of social media platforms on the use of their APIs in an attempt to monetize their data. Embarking on large-scale data collection initiatives is becoming more and more difficult, because the number of messages that can be downloaded is limited (e.g., Twitter/X restricts this to 500 messages per day), access to the APIs requires a fee, or the use of the data for training machine learning or AI models is explicitly forbidden (e.g., Reddit).

As already stated, the expansion of datasets and lexicons across various geographical and linguistic landscapes is on the rise. However, the absence of standardized protocols and guidelines for data collection hinders the advancement of robust multi- and cross-lingual NLP tools. Effective data collection methodologies play a pivotal role in this process. Approaches reliant on keywords are susceptible to introducing biases centered around topics or authors, as highlighted by Wiegand et al., (2018). Societal, ethnographic and community-centric aspects are often overlooked and underrepresented, thereby limiting the identification of hate speech to mere advanced pattern recognition. Different social groups may uphold distinct standards regarding what constitutes hate. For instance, in the drag community the use of the b-word is often used as a term of affection/ endearment rather than in its derogative meaning. Neglecting to include or accurately represent these diverse perspectives within the data can lead to detrimental consequences. Finally, multimodal datasets for hate speech are still rare (Hosseinmardi et al., 2015; Zhong et al., 2015; Suryawanshi et al., 2020), with the written-only modality being predominant, disregarding the potential synergies offered by multiple modalities like images and text – also in the light of the fact that the web is increasingly becoming a visual medium.

Data annotation

Availability of annotated data is key to any supervised machine learning approach. Nevertheless, the availability of annotated data also plays a role for more qualitative studies, making it easier to compare related, if not the same, language phenomena. Here we will focus on reporting on practices that have been developed within NLP.

Early work on hate speech adopted a holistic approach by annotating a message as either fulfilling the specific definition that was adopted or not. More fine-grained annotations have been developed following the proposal of a multi-axis annotation approach by Waseem et al., (2017). On the other hand, a differentiated approach has been implemented for the OffensEval 2019 and 2020 shared tasks. In this case, the task organizers developed a three-layer annotation scheme: the first layer is responsible for the binary distinction between offensive and non-offensive messages; the second layer further distinguishes whether the offensive message contains a target or not; and lastly the third layer specifies whether the target is an individual, a group or another entity. The proposed solution has a further advantage that by combining all layers together it becomes possible to further identify more heinous language phenomena such as abusive language or hate speech.

The annotation of hate speech usually takes place in isolation: in no available datasets do the annotators (or the machines) have access to the (full) context of the occurrence of the message under analysis. While main posts (i.e., messages potentially initiating a thread or a discussion with other users) can be annotated quite reliably as isolated messages, in the case of replies, the lack of context has a potentially devastating impact: a smiley face emoji as a reply to a hateful message is also loaded with hate. Annotation in isolation not only affects the context of the occurrence of the message, but it also comprises the lack of any socio-demographic information on the user who posted a message. Having access to this data is not always possible, and even when it is, it may not be desirable due to privacy and other ethical considerations. At the same time, such data could help annotators to avoid biased annotations when it comes to specific slurs known to be re-appropriated by minorities.

As Artstein and Poesio (2008) have pointed out, there is no correlation between inter-coder agreement and system performance, however it is undeniable that the annotation of hate speech is a difficult task. The current datasets present, in general, medium levels of agreement among their coders. Furthermore, potential additional sources of disagreement can be prompted by the formulation of the annotation guidelines. Röttger et al., (2022) discuss two major paradigms for the development of annotation guidelines for highly subjective tasks. The first, referred to as the descriptive paradigm, does not impose many constraints on the annotators' subjectivity, thus allowing the annotations to capture many beliefs. The approach allows for a more granular modelization of the annotations, shedding lights on those cases that are mostly subjective and thus representing informative cases of disagreement. In the case of hate speech, it is easy to see how this annotation paradigm can support the identification of cases where different annotators perceive different levels of hate. The second paradigm, referred to as the prescriptive paradigm, on the other hand, requires annotators to strictly follow the provided annotation guidelines, aiming to minimize subjective interpretations. Some of the issues affecting the annotation of existing datasets can be brought back to a lack of clear alignment of the developed annotation guidelines with any of the two paradigms.

Being a highly subjective task, the annotation of hate speech has always seen multiple annotators expressing their judgements on each message, regardless of whether the annotations are conducted by experts or by crowd workers. Disagreements in the annotations are usually disregarded by either collapsing them into a single label based on the majority of the annotation preferences, or resolved by asking the annotators to resolve them. Only recently has there been an increased awareness of considering disagreements – in this kind of task – as additional sources of information rather than noise or errors (Plank et al., 2014; Leonardelli et al., 2021; Cabitza et al., 2023).

Ethical application of the AI technologies by social media companies

Zeynep Özarslan

Social media platforms provide milieus and tools that facilitate freedom of expression. Yet, as Ullmann and Tomalin (2020) observe, these platforms have "created new forms of swift and efficient communication in which hate speech can be expressed almost instantaneously online, and often anonymously." (p. 70). Furthermore, the pervasive use of AI technologies, particularly generative AI, has positioned social media platforms at the forefront of mediating and amplifying new forms of bias, discrimination and hatred. Hate speech proliferates on social media through both overt and covert tactics; such as memes, GIFs, emojis, filters, the use of fake identities and influencers promoting racism. Despite the efforts of social media and AI technology companies to prevent hate speech through policies and content moderation processes, these technologies are paradoxically employed to perpetuate hatred and discrimination, thereby reshaping oppression along the lines of race, gender, sexuality, language, religion, etc.

As Loebbecke et al., (2021) argue, since platforms operate globally in culturally diverse markets, their terms of service must encompass various definitions of hate speech to satisfy a global audience. Rather than a universally applicable set of rules, diverse regulations worldwide require social media platforms to devise processes for the removal of illegal content, such as hate speech. To fulfil this mandate, platforms invest in content moderation systems to address such malicious content (Loebbecke et al., 2021). Typically, human moderators review offensive messages on social media platforms upon receipt of user complaints. If the content is classified as hate speech, it is removed from the platform; otherwise, it remains accessible. Nonetheless, even when offensive content is deleted, the damage to the recipient may persist due to prior exposure. Moreover, the sheer volume and velocity of online hate speech content can overwhelm human moderators. Consequently, "in the last few years, an emerging generation of automatic hate speech detection systems has started to offer new strategies for dealing with this particular kind of offensive online material." (Ullmann & Tomalin, 2020, p. 69). Additionally, AI technologies such as natural language processing (NLP), machine learning (ML), deep learning and explainable artificial intelligence (XAI) have been employed to detect hate speech and abusive language by extracting text-based, user-based, and network-based features, as well as by identifying online bullies and hate-related keywords (Mehta & Passi, 2022). "However, lacking a mutual understanding of what constitutes hate speech complicates the endeavor. Different definitions, regulations, and contexts challenge the development and deployment of the related AI systems." (Loebbecke et al., 2021).

While AI technologies and tools are valuable for detecting online hate speech, their ethical application is imperative. A corpus of AI ethical guidelines and frameworks has been recently formulated to mitigate the potential harms of new AI technologies. Scholars have analyzed and synthesized these principles and guidelines to articulate a coherent vision of "ethical AI" and the ethical and technical standards necessary for its operationalization. For instance, Royakkers et al., (2018) identify six recurring themes emerging from their analysis; privacy, autonomy, security, human dignity, justice and balance of power. They note a lack of adequate supervision in areas such as discrimination, autonomy, human dignity and unequal balance of power. Likewise, Jobin et al. (2019) examined 84 ethical AI documents and identified eleven ethical values and principles that emerged from their content analysis: Transparency, justice and fairness, non-maleficence, responsibility, privacy, beneficence, freedom and autonomy, trust, dignity, sustainability and solidarity, with an emerging convergence around transparency, justice and fairness, non-maleficence, responsibility and privacy. Correspondingly, Morley et al. (2020) derived an ethically aligned ML definition that encompasses being "(a) beneficial to, and respectful of, people and the environment (beneficence); (b) robust and secure (non-maleficence); (c) respectful of human values (autonomy); (d) fair (justice); and (e) explainable, accountable, and understandable (explicability)." (p. 2145). Similarly, Hagendorff (2020), analyses major AI ethics guidelines and recommendations and highlights recurring themes such as accountability, privacy or fairness. Notably, Hagendorff (2020) points out that substantial technical efforts have been made to achieve ethical benchmarks in accountability, explainable AI, fairness and discrimination aware data mining, as well as privacy, yet highlights that "hardly any guideline discusses the possibility for political abuse of AI systems in the context of automated propaganda, bots, fake news, deep fakes, micro targeting, election fraud, and the like … [and] the issue of a lack in diversity within the AI community." (Hagendorff 2020, p. 103, 105). Furthermore, Hagendorff (2020) contends that AI has significantly contributed to diminishing social cohesion and fostering radicalization, the decline of rational public discourse and social divides (p. 110). On the other hand, Robles Carrillo (2020) argues that despite the presence of basic principles and common elements, ethical conceptions and principles are subject to variation across traditions, cultures, ideologies, systems and countries, and the substance of "the ethical" evolves with changing times and societies (p. 3).

The use of AI systems and tools can be unethical, depending on how corporations employ the technology. Nonetheless, strategies and techniques are available to prevent such unethical practices. Social media platforms should prioritize algorithmic justice, fairness and equity when developing and using AI systems or tools, referring to the elimination of bias and discrimination risks within datasets by acquiring and processing accurate, complete and diverse data, particularly during the training phase. Social media platforms should develop and/or use non-maleficent AI systems, which refers to the avoidance of potential harms, such as discrimination and violation of privacy. In other words, corporations should be aware of the negative impacts of AI systems and take steps to mitigate them. Correspondingly, Zhuo et al., (2023) maintain that when the training data includes biased representations of specific groups of individuals, the large language models (LLM) provide predictions that are unfair or discriminatory towards those groups. Therefore, to prevent this, "it is essential to ensure that the training data is diverse and representative of the population for which it will be used, ...and to actively discover and eradicate any potential biases in the data." (Zhuo et al., 2023). Exclusionary norms are another ethical consideration, implying that the training data represents only a segment of the population, such as a single culture. This could lead to the model's inability to understand or produce content for underrepresented groups, such as speakers of different languages or people from various cultures (Zhuo et al., 2023). Consequently, AI systems and tools must be diverse and inclusive, benefiting as wide an audience as possible.

Monolingual bias, where AI models are trained exclusively on data in one language, can prevent the models from comprehending or generating text in other languages, thereby denying benefits to non-speakers and potentially leading to biased or unfair predictions about those groups. To counter this, it is imperative to ensure that the training data encompasses a substantial proportion of diverse, high-quality corpora from various languages and cultures (Zhuo et al., 2023). The authors of this same study also highlight potential toxicity in LLMs, referring to the models' capacity to generate or understand harmful or offensive content. Toxicity may stem from training data containing offensive language, indicating that the model will recognize and generate offensive content during user interactions. To mitigate this risk, it is essential to ensure that the training data is devoid of offensive language and to proactively remove any offensive material that may be present (Zhuo et al., 2023).

Reliability is another ethical consideration in AI systems. To develop a reliable AI system, the training data should not include any false, inaccurate or misleading information. Otherwise, misleading outputs will be generated by the AI system. Therefore, it is crucial to keep the training data up-to-date and continuously monitor and update the AI systems to ensure they give the most accurate information. In addition, Buolamwini and Gebru (2018) examine the ethical concerns in computer vision technologies, specifically automated facial image analysis, underlining the issues of algorithmic fairness, accuracy, transparency and accountability. Likewise, Gebru et al., (2018) state that machine learning models can reproduce or amplify unwanted societal biases reflected in training datasets. Therefore, they propose datasheets for datasets that, "have the potential to increase transparency and accountability within the machine learning community, mitigate unwanted societal biases in machine learning models, facilitate greater reproducibility of machine learning results, and help researchers and practitioners to select more appropriate datasets for their chosen tasks." (Gebru et al., 2018).

In this regard, how can the ethical use of AI help social media platforms prevent the generation and dissemination of hate speech? To address the issue effectively, a multi-layered approach could be applied as follows:

Diverse and representative training data: Inclusive use of multiple languages, dialects, cultures and demographic representations in training datasets could help mitigate biases that lead to discrimination or hate speech. In other words, incorporating diverse linguistic and cultural backgrounds in training data could reduce monolingual biases, thereby helping social media algorithms detect hate speech effectively. Furthermore, establishing protocols to detect discriminatory patterns across demographics could prevent the reproduction of systemic biases on social media platforms.

Analyzing contextual cues: When identifying and flagging potential toxicity, platforms could develop AI systems to analyze contextual cues considering a user's historical engagements and activities and other interactions within the network. Therefore, developing such AI systems for contextual analysis could help identify hate speech more precisely, thus aiding in its prevention and dissemination.

Fairness and transparency standards: It is essential for social media platforms to be transparent about their AI usage, providing explanations regarding the types of training data, content moderation and decision-making processes employed by their algorithms. Establishing transparency standards in using AI systems and regularly publishing reports could help develop more effective methods to detect hate speech on social media platforms.

Compliance with AI ethical standards: Social media platforms should adopt up-to-date AI ethical guidelines and frameworks ensuring transparency, justice, non-maleficence, responsibility and privacy in their algorithms. In addition, forming independent entities from a diverse array of stakeholders – including ethicists, sociologists, legal experts, technologists, user representatives, non-governmental organizations (NGOs) and civil society groups – to conduct regular audits of social media algorithms could help develop to identify unethical practices and to evaluate the efficacy of hate speech filtering mechanisms. This interdisciplinary approach ensures a comprehensive consideration of ethical concerns. Therefore, these independent entities could provide recommendations for the customization of systems so that they comply with diverse ethical conceptions and offer insights into preventing hate speech.

Empowering human moderators and users with AI: Ethical AI systems on social media platforms could help forecast and flag potential hate speech content, thereby helping prevent the dissemination of hate speech. This could also reduce the workload of human moderators and improve moderation accuracy. Moreover, establishing AI-powered crisis management protocols for quick responses to surges in hate speech could help prevent the spread of hate speech. Furthermore, developing AI-assisted filters could empower users to report hate speech and manage content exposure, enabling proactive content moderation.

Cross-platform collaboration: Deployment of AI ethical guidelines and frameworks is inherently complex and multifaceted. Therefore, to establish agreements among different platforms for a cohesive global strategy against hate speech and to form independent committees across different platforms to share best practices in ethical AI usage and coordinate efforts against hate speech, i.e. cross-platform collaboration, could be beneficial for the sharing of insights and strategies to combat hate speech effectively.

Educational campaigns: Social media platforms could launch educational campaigns on the ethical use of AI in detecting and eliminating hate speech. In addition, ethical AI systems could be utilized to promote positive and educational counter-narratives that weaken the impact of hate speech, which can lead to more positive and constructive engagement.

COUNTERING HATE SPEECH: BEYOND POLICIES AND REGULATIONS

Although state regulations and policies of social media platforms prohibiting and even penalising hate speech aim to combat discrimination, these efforts must always be balanced against freedom of expression, as noted in the previous section. Because the enforcement of these regulations can sometimes pose a threat to freedom of expression. In some cases, the line between harmful discourse and legitimate dissent is blurred, leading to concerns that restrictions may be overreaching and stifling open debate and critical thinking. Moreover, in the hands of totalitarian regimes, such laws can become a tool to silence opposition. For example in Turkey's case, laws such as "incitement to hatred" have been criticized for being applied in ways that suppress dissent rather than protect vulnerable groups. This highlights the potential for such regulations to be misused, particularly in politically charged environments, where the focus shifts from combating discrimination to targeting opposition voices.

On social media platforms, even when there are clear community guidelines aimed at banning hate speech, the challenges of identifying and categorizing such content lead to inconsistent enforcement, as discussed in great detail above. Due to the inherent difficulties in identifying nuanced or covert forms of hate speech, much of this content remains online, allowing it to proliferate and further reinforce harmful narratives. As a result, while these platform policies are functional, they do not prevent the spread of hate speech.

It is crucial to recognize that the core issue lies in defining hate speech without infringing on freedom of expression. Over-regulation and heavy-handed enforcement can create an environment of fear and self-censorship, where people may hesitate to voice legitimate grievances or engage in critical discussions.

Furthermore, these regulations and prohibitions may fall short of providing a lasting solution to the root causes of social prejudice. Simply banning hate speech does not address the underlying factors that give rise to hate speech, such as systemic inequality. Relying solely on regulations to foster an inclusive environment is insufficient and, as evidenced by examples from authoritarian contexts, potentially dangerous.

We emphasize the importance of developing the technological tools discussed in this report to identify issues, generate data on the issue, and monitor the discrimination that can evolve in society based on context. However, we also believe that addressing the root causes of the problem is essential when proposing solutions. Freedom of speech must remain a cornerstone of any initiative aimed at addressing hate speech, as fostering inclusivity cannot come at the expense of silencing dissent.

In this section, various approaches and initiatives aimed at preventing the production and dissemination of hate speech, as well as reducing its potential impact on targeted groups will be explored.

First of all, Susan Benesch will examine the concept of counter speech as a potential solution to the problem of hate speech. Subsequently, a number of case studies will be presented, illustrating the potential of social media campaigns. Finally, the crucial role of education and the necessity of engaging with children and youth will be emphasized. In this context, the contributions of civil society organizations are of paramount importance, as they can significantly influence positive change, play a vital role in fostering dialogue and promoting inclusivity through education and awareness-raising efforts. It is also important to acknowledge that accelerating the implementation of these solutions is essential for addressing the underlying causes of hate speech and discrimination.

Counterspeech approaches to undermining hate speech

Susan Benesch

Amid great concern over what to do about hate speech, including online, some countries (most recently Scotland, at this writing) have increased legal penalties for such speech, but speech laws can threaten freedom of expression and are often used to silence minority or opposition voices. There is another option: grassroots efforts to improve online discourse. Thousands of internet users regularly respond directly to hate speech, to refute or undermine it using a variety of communicative strategies.

Some offer factual information to correct hateful disinformation and prevent other readers from being persuaded by it. Others try to educate internet users about hate speech by posting in a larger forum where more users will view it. Still others use humor to help their responses gain an audience (people may follow an account because it is funny, for example) or to decrease the emotional burden of responding to such disturbing content. Finally, some try to empathize with purveyors of hate speech. Empathy can convey a sense of understanding that might lead to the original hateful speaker changing their behavior or beliefs (although this can be difficult to accomplish). Many of those who engage in counterspeech go about it alone, while others form groups to coordinate responses and support each other.

Counterspeech is not a new concept, although discussions and the study of it have become more common in recent years. In the United States, the concept of counterspeech is often traced back to U.S. Supreme Court Justice Louis D. Brandeis, who in the 1927 case Whitney v. California, proposed the idea that replying to harmful speech, not censoring it, is the best policy. In a decision upholding the conviction of a California woman who had worked to establish the Communist Labor Party of America, Brandeis declared:

"If there be time to expose through discussion the falsehood and fallacies, to avert the evil by the processes of education, the remedy to be applied is more speech, not enforced silence."

U.S. lawyers often call this the counterspeech doctrine, though Brandeis himself never used the term.

There is a growing body of scholarship on the topic of counterspeech, with many of these studies focusing on whether or not counterspeech is effective. Even though many researchers have tried to answer this question, there are considerable challenges, including how to define effectiveness. Counterspeech can have a positive effect on discourse in several ways. It can convince people to stop posting harmful speech, by changing their beliefs or only their behavior. (The latter is possible since people can come to fear criticism or social sanction for publicly expressing a belief, even if they still hold it.) Discourse may also improve without any change in the views or online expression of people posting hatred. Instead, counterspeakers can succeed by influencing the "audience" – the people who read their comments. For those who may already agree with the views shared by counterspeakers, but do not yet feel brave enough to share their own, counterspeech can encourage them to chime in, thus gradually shifting discourse toward the views expressed in counterspeech, even if no beliefs change. Other members of the audience may not have formed a solid opinion about the topic being discussed yet, and counterspeech could help prevent them from believing or sharing harmful misinformation.

Counterspeech is different from counternarrative, a broader strategy that involves challenging dominant narratives or ideologies that perpetuate injustice, discrimination, or inequality. These narratives generally include an alternative interpretation, analysis or understanding of historical events, social issues, cultural norms, or political ideologies.

While counterspeech and counternarrative share similar objectives in challenging harmful ideologies, they differ in their scope, focus, and methodology. Counterspeech primarily addresses individual instances of hate speech, aiming to refute or undermine specific arguments or messages. In contrast, counternarrative campaigns aim to deconstruct and subvert the overarching narratives that sustain systemic oppression or marginalization.

Countering hate speech through activism: Two successful examples

İlayda Ece Ova

Institutional and legal efforts at both the national and INGO (International Non-Governmental Organization) levels to counter hate speech are highly visible due to their extensive resources and outreach capabilities. The UN has adopted a dedicated agenda on combating hate speech, designating June 18 as the International Day for Countering Hate Speech, building on its 2019 UN Strategy and Plan of Action on Hate Speech. Meanwhile, the EU funds numerous projects across sectors and countries through various bodies, providing consistent financial support for civil society organizations working in this field. **The High Level Group on combating hate speech and hate crime**[33] is the platform where the EU's agenda on hate speech is set, while **the CERV program**[34] provides the financial framework that aims to protect and promote EU rights and values that align with anti-hate speech work. Notable examples include projects and programs such as **Combating Hate Speech in Sport**[35], a project that aims to tackle hate speech in sports by offering technical support to public authorities and sport stakeholders; **Combating Anti-LGBTIQ Violence and Hate Speech**[36], a project that seeks to prevent and address hateful and intolerant discourse, violence and discrimination based on sexual orientation, gender identity or expression; **The Stand-Up Project**[37], an inter-institutional model to improve cooperation between different organizations in the fight against hate

[33] **European Commission.** (2023). *High-Level Expert Group on Fake News and Online Disinformation*. European Commission Transparency Register. https://ec.europa.eu/transparency/expert-groups-register/screen/expert-groups/consult?lang=en&groupID=3425

[34] **European Commission.** (2024). *Citizens, Equality, Rights and Values programme (CERV)*. European Commission Funding & Tenders Portal. https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/programmes/cerv

[35] **Council of Europe.** (2024). *Combating hate speech in sport: About the project*. Council of Europe. https://pjp-eu.coe.int/en/web/combating-hate-speech-in-sport/about-the-project

[36] **Council of Europe.** (2024). *Combating anti-LGBTIQ+ violence*. Council of Europe. https://www.coe.int/en/web/sogi/combating-anti-lgbtiq-violence

[37] **STAND-UP Project.** (2024). *STAND-UP: Fighting hate in the EU*. STAND-UP Project. https://stand-up-project.eu/

crimes; and **Facts Against Hate**[38], an international and cross-institutional collaboration that developed reporting tools and referral mechanisms. These projects, although varying in aims and outcomes, focus on multi-sector collaboration, data collection improvements, digital tools, support for affected individuals and communities, and targeted efforts to combat discrimination in areas such as sports and online spaces, as well as within marginalized communities.

In the U.S., major organizations combat hate speech through advocacy, education, legal action, and community engagement. **The Anti-Defamation League**[39] (ADL) addresses extremism and antisemitism with initiatives like **Stop Hate for Profit**[40], while the **Southern Poverty Law Center**[41] (SPLC) tracks hate groups and raises awareness through its **Hate Map**[42]. **The Human Rights Campaign**[43] (HRC) focuses on hate speech targeting LGBTQ+ individuals, and the **American Civil Liberties Union**[44] (ACLU) works to defend free speech while mitigating its harmful effects. **The Center for Countering Digital Hate**[45] (CCDH) tackles online hate and disinformation, while **Faith in Action**[46] fosters community resilience through interfaith collaboration. The **NAACP**[47] (National Association for the Advancement for Colored People) combats racial hate and discrimination, and **CAIR**[48] (Council on American-Islamic Relations) addresses Islamophobia. **GLAAD**[49] (Gay

[38] **European Union Agency for Fundamental Rights (FRA).** (2024). *Facts against hate*. FRA. https://fra.europa.eu/en/promising-practices/facts-against-hate-0

[39] **Anti-Defamation League (ADL).** (2024). *Anti-Defamation League*. ADL. https://www.adl.org/

[40] **Anti-Defamation League (ADL).** (2024). *Stop hate for profit*. ADL. https://www.adl.org/stop-hate-profit-0

[41] **Southern Poverty Law Center (SPLC).** (2024). *Southern Poverty Law Center*. SPLC. https://www.splcenter.org/

[42] **Southern Poverty Law Center (SPLC).** (2024). *Hate map*. SPLC. https://www.splcenter.org/hate-map/

[43] **Human Rights Campaign (HRC).** (2021). *Human Rights Campaign*. HRC. https://www.hrc.org/

[44] **American Civil Liberties Union (ACLU).** (2024). *American Civil Liberties Union*. ACLU. https://www.aclu.org/

[45] **Center for Countering Digital Hate (CCDH).** (2022). *Center for Countering Digital Hate*. CCDH. https://counterhate.com/

[46] **Faith in Action**. (2024). *Faith in Action*. Faith in Action. https://faithinaction.org/

[47] **National Association for the Advancement of Colored People (NAACP).** (2024). *NAACP*. NAACP. https://naacp.org/

[48] **Council on American-Islamic Relations (CAIR).** (2024). *Council on American-Islamic Relations*. CAIR. https://www.cair.com/

[49] **GLAAD.** (2024). *GLAAD*. GLAAD. https://glaad.org/

& Lesbian Alliance Against Defamation) monitors anti-LGBTQ+ media portrayals and promotes positive narratives, while **Not In Our Town**[50] (NIOT) empowers local communities to take grassroots action against hate, creating a multifaceted national response.

The widespread adoption of anti-hate speech initiatives by national and international organizations has significantly increased visibility and contributed to the dominance of institutional and legal solutions. However, this dominance may lead to the perception that stricter legal regulations and large donor-funded programs are the ultimate solutions. While beneficial, the rise of right-wing and authoritarian regimes worldwide highlights the vulnerability of state-led efforts and how easily inclusive policies can be reversed. As state-backed, inclusive initiatives decline under the influence of right-wing, misogynistic and anti-LGBTI+ ideologies, activism against hate speech becomes increasingly critical and vital. This underscores the importance of civil society and citizen-led responses. Long-term societal change requires context-specific, localized and culturally relevant activism driven by civil society organizations, grassroots movements, and dedicated individuals.

Grassroots activism against hate speech exemplifies how context-specific actions can address unique challenges while facilitating knowledge exchange and effective practices across international networks. Anti-hate speech campaigns share common strategies, including raising awareness, promoting tolerance and building inclusive societies. They focus on educational initiatives, community engagement, and youth mobilization to foster long-term resilience. Social media plays a key role in spreading counter-narratives, with many campaigns emphasizing non-confrontational responses to de-escalate online conflicts. Fact-checking and combating misinformation are essential, given the significant role of disinformation in spreading hate. Campaigns also provide support for victims through psychological, legal, and online resources, while advocating for improved platform moderation and policies. Many use culturally resonant symbols to convey simple yet impactful messages, while partnerships with governments, NGOs and the private sector help influence policies and strengthen content moderation. With global coordination and local adaptability, these efforts effectively combat hate speech across diverse contexts.

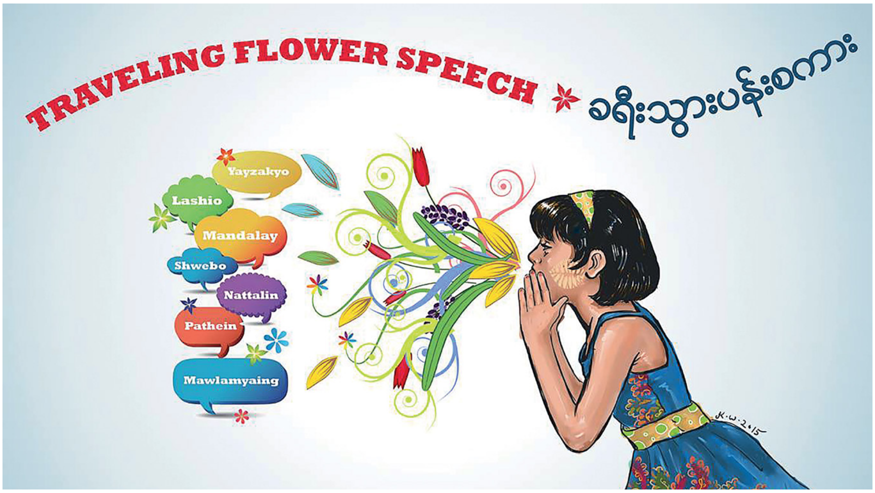
To illustrate the impact of such approaches, two best-practice examples of campaigns that have effectively challenged hate speech through innovative, culturally relevant strategies will be highlighted.

[50] **Not In Our Town (NIOT).** (2024). *Not In Our Town*. NIOT. https://www.niot.org/

Panzagar (Flower Speech) Campaign

The *Panzagar* ("flower speech") Campaign[51], launched in Myanmar in 2014 by Burmese youth leader Nay Phone Latt, was a response to escalating anti-Muslim violence. The campaign adopted the symbolic use of flowers, which represent peace in Myanmar. Civil society reports highlighted that online hate speech in the country was primarily directed toward Muslims. The campaign's core strategy involved collaborating with young local graphic designers to create anime-inspired visuals featuring characters with flowers emerging from their mouths, which were widely circulated on social media.[52]

Figure 2: Initial Panzagar campaign imagery[53]



In 2014, Facebook introduced a user reporting mechanism for abuse in Myanmar, but the country's unique demographic and legal context required additional targeted actions. At the time, Myanmar had one of the world's lowest rates of internet and mobile phone usage, and government censorship was pervasive. As a result, the campaign had both online and offline components and drew inspiration from Buddhist organizations that countered the portrayal of Muslims as a threat to the Buddhist majority.

[51] **Beautiful Trouble.** (2024). *Flower speech campaign*. Beautiful Trouble Toolbox. https://beautifultrouble.org/toolbox/tool/flower-speech-campaign

[52] **Panzagar.** (2014). *Panzagar Facebook page*. Facebook. https://www.facebook.com/panzagar

[53] **Support Flower Speech.** (2014). *Support Flower Speech Facebook page*. Facebook. https://www.facebook.com/supportflowerspeech/

Activist Thinzar Shunlei Yi reflects on the campaign's strengths and weaknesses, highlighting how the #FlowerSpeech campaign successfully encouraged widespread participation through social media, public events, music, and stickers, turning passive supporters into active participants. The campaign referenced the Buddhist principle of "right speech", promoting ethical and non-abusive communication. Its success was largely due to leveraging a culturally significant image – Myanmar's national flower, the padauk – held in people's mouths, ensuring that the message was easily understood. However, early criticisms included concerns about the sexualization of initial graphics, the lack of Burmese-specific imagery and stereotyping. The campaign responded to these criticisms by adjusting its visuals. Another critique pointed out that holding a flower in one's mouth could be interpreted as promoting silence instead of active resistance to hate. Despite these limitations, the campaign provided effective tools to diffuse hate speech and promote constructive dialogue.

#IAmHere International Campaign

The **#IAmHere** movement originated in Sweden as #JagArHar, founded by Iranian-born journalist Mina Dennert.[54] Dennert launched the initiative after observing a surge in hateful content on social media and sought to counter it with calm, non-confrontational responses. The campaign gained media attention and rapidly grew to 75,000 members in Sweden, inspiring similar groups in Italy, France, Slovakia, Poland, and the UK. Today, fourteen #IAmHere groups operate globally, sharing the same mission of countering online hate speech by promoting fact-based, non-confrontational dialogue, supporting those targeted by hate, and advocating for healthier digital spaces.

The international campaign focuses on fact-checking and preventing the spread of disinformation. Volunteers mobilize to counter hateful comments on social media, support individuals targeted by hate speech, and advocate for legal regulations.[55] Disinformation in the digital sphere is often spread through coordinated networks of trolls and bots, which amplify harmful narratives and manipulate online discussions. Troll groups operate with military-like hierarchies, using structured and strategic coordination to dominate conversations, suppress mod-

[54] **BBC News.** (2019, June 10). *'#IAmHere': The people trying to make Facebook a nicer place.* BBC News. https://www.bbc.com/news/blogs-trending-48462190

[55] **#IamHere International.** (2024). *About us.* #IamHere International. ttps://iamhereinternational.com/about-us/

erate voices, and aggressively spread disinformation. To push back against these tactics, the #IAmHere movement adopts a similarly organized approach, mobilizing its members in a focused and strategic manner to challenge hate speech and misinformation. The campaign’s final step, “pushing hate down,” involves advocating for algorithm and platform regulations that reduce the visibility of hateful comments. However, as discussed earlier, relying solely on social media platforms to improve policies and algorithms presents challenges. Instead, the campaign’s emphasis on freedom of speech and constructive, fact-checked responses provides a vital model for grassroots and civil society-led anti-hate speech efforts globally.

Alternative approaches to fostering positive engagement

Peter Adams, Metin V. Bayrak, Alex Mahadevan

Leveraging influencers for inclusive narratives

Popular content creators may be flagged as vectors of misinformation and hate speech, but on the other hand, the trust and scale influencers and celebrity content creators enjoy with children and teenagers also holds great opportunity in the development of digital media literacy and hate speech countermeasures. The same way researchers are developing artificial intelligence-based tools to combat generative AI disinformation, so too can NGOs, news organizations and other civic organizations fight falsehoods by making use of traditional vectors of misinformation themselves. Influencers and content creators thrive on social media platforms like YouTube, TikTok, Instagram and Twitch, so social media companies have a stake in ensuring their talent does not spread misinformation or hate speech. Platforms such as Facebook, Instagram (Meta) and TikTok collaborate with fact-checkers, and all platforms have community guidelines addressing hate speech and discrimination, as discussed in previous sections of this document. However, there is significant and well-founded criticism regarding the effectiveness of these guidelines. Additionally, Meta discontinued its fact-checking program in January 2025, raising further concerns about the reliability of content moderation on its platforms. However, one successful example worth highlighting is the case where **The Australian Associated Press worked with TikTok creators**[56] to develop interventions on the platform aimed at debunking misinformation and building media literacy skills.

Another strength influencers and content creators bring is their ability to create native content that does not "feel" like an advertisement or a public service announcement. They are also experts on how to craft videos and graphics with the highest potential to "go viral". Since 2018, MediaWise, the media literacy initiative of the Poynter Institute for Media Studies, has also run an Ambas-

[56] **Mediaweek.** (2024, February 28). TikTok and AAP partner to empower creators to fight misinformation. Mediaweek. https://www.mediaweek.com.au/tiktok-and-aap-partner-to-empower-creators-to-fight-misinformation/

sador Program made up of well-known journalists, along with YouTubers, authors, beauty and engineering influencers, LGBTQ+ focused content creators and athletes. For example, Guatemalan badminton champion and Olympian Kevin Cordon led a series of media literacy videos[57] as part of a course delivered on WhatsApp.

Regarding efficacy, a **Stanford University study** found that courses centered on U.S. journalist and aging-issues influencer Joan Lunden increased the discernment of misleading headlines.[58]

Finally, the National Association for Media Literacy Education contributed the curriculum for YouTube's **Hit Pause**[59] campaign, which collaborated with influencers such as **children's TV star Blippy**[60]. Each video drew at least 40 million views with some reaching more than 200 million.

In light of the presented examples, it is possible to propose a number of strategies through which civil society organizations may leverage the potential of social media to reduce hate speech:

- **Storytelling Campaigns:** Influencers can share personal stories or narratives of individuals affected by hate speech and misinformation, humanizing the issue and promoting empathy. The success of such campaigns can be measured by engagement metrics and sentiment analysis in the comments and shares.
- **Educational Series:** Collaborate with influencers to create a series of educational content that breaks down complex topics related to hate speech and misinformation. Success can be gauged through viewership numbers, completion rates of the series and pre- and post-campaign surveys to assess changes in understanding and attitudes.

[57] **Poynter Institute.** (2024). *MediaWise in Guatemala*. Poynter Institute. https://www.poynter.org/mediawise/international/guatemala/

[58] **Dyakon, T.** (2020, December 10). *Poynter's MediaWise training significantly increases people's ability to detect disinformation, new Stanford study finds.* Poynter. https://www.poynter.org/news-release/2020/poynters-mediawise-training-significantly-increases-peoples-ability-to-detect-disinformation-new-stanford-study-finds/

[59] **Hit Pause.** (2023, Feb 7). *Digital Wellbeing for Families* [Video]. YouTube. https://www.youtube.com/playlist?list=PL4SOO4mxq3nsE6nzCV-QDGzg6VF7cFzGh

[60] **Hit Pause.** (2023, Feb 7). *Hit Pause with a silly stretch break* [Video]. YouTube. https://www.youtube.com/watch?v=xdKDcKuRs3M

- **Interactive Challenges:** Encourage influencers to launch challenges that promote positive messaging or debunk myths, using interactive content like quizzes and response videos. Metrics for success include participation rates, the virality of content (shares and likes) and the quality of discourse in the comments.
- **Partnerships for Policy Advocacy:** Influencers can be used to advocate for policies against hate speech and misinformation. Success here can be measured by the number of signatures on petitions, attendance at related events and policy changes influenced.
- **Digital Literacy Workshops:** Host live sessions with influencers where audiences can learn about identifying and responding to misinformation and hate speech. Effectiveness can be measured by attendance rates, interaction during sessions and feedback forms.

It is important to note that working with influencers can be risky, as you cannot control what they do or say outside of the specific program. Furthermore, streamers specifically, may have spent years creating thousands of hours of content some of which could be considered questionable. Before seeking collaboration, considerable time must be spent vetting the previous work of influencers, content creators or celebrities to ensure they have not produced polarizing, biased or polarizing political content. It is also best practice to work in coordination with the influencer on content development and carefully edit the material to keep it on message and free from bias. It helps to provide a script outline – or even the full script itself – to the content creator ahead of time.

Building inclusivity: Educational approaches for youth

The spread of hate speech, disinformation and extremist rhetoric not only perpetuates intolerance, hate and violence, it also corrupts productive civic discourse and undermines the development of key democratic mentalities among all, but in the context of this section we will especially focus on children and teens.

Disinformation – and especially conspiratorial content and ideas – is one of the primary vectors for the spread of hate speech and extremist ideologies. Far right communities are highly active online and often target teens using common disinformation tactics and conspiratorial tropes, including the exploitation of cognitive biases, the use of logical fallacies, the fabrication of evidence and tricks of context meant to elicit a strong emotional reaction. Like other disinformation purveyors, extremists capitalize on universal human needs, such as the desire for understanding, connection, community, purpose and agency (these are sometimes conceived of as "push and pull" factors[61]). These needs are particularly acute among teens and young adults, who tend to be more focused on social relationships and potential social rewards.

It is vital that parents, guardians, educators and other stakeholders in the lives of young people prepare them to recognize and resist extremist messages and other kinds of harmful disinformation. These influences make their way into virtually every aspect of teens' information streams. They show up as posts and comments on mainstream social media platforms, often using evasive, coded or sanitized language and symbols; in closed messaging groups; on fringe message boards and social sharing sites; through white nationalist music; and on interest-based platforms like Discord and game-streaming and live chat platforms like Twitch.

Empowering adults to address young people's awareness of hate speech

While fostering awareness among children and youth is essential, adults – specifically parents and educators – also play a crucial role in shaping inclusive environments that counter hate speech. Their ability to address discrimination, encourage critical thinking and create safe spaces significantly impacts how young people engage with and respond to harmful narratives.

Addressing hate speech aligns with educational objectives like fostering empathy and tolerance. Adults who understand these values can integrate them into their educational practices, enriching children's learning experiences.

To support children in navigating and resisting hate speech, adults need to be equipped with the right tools and knowledge. Some key areas where adult education can strengthen efforts to counter discrimination include:

- **Promoting Positive Social Behavior:** Educating adults on addressing hate speech helps create environments where children learn respect and appreciate diversity, fostering positive behavior from a young age.
- **Protecting Mental Health:** Hate speech can harm children's mental well-being, leading to anxiety, depression and low self-esteem. By equipping adults to recognize and combat it, we help protect children's mental health.

[61] **TED.** (2017,Sep 17). *How Young People Join Violent Extremist Groups – and How to Stop Them* [Video]. YouTube. https://www.youtube.com/watch?v=HY71088saG4

- **Building Inclusive Communities:** Hate speech fosters division and exclusion. Empowering adults with the skills to confront it promotes more inclusive and cohesive communities where everyone feels valued.
- **Promoting Critical Thinking:** Adults trained to understand the dangers of hate speech can teach children critical thinking. Encouraging kids to question and analyze hateful messages helps them develop resilience against harmful ideologies.
- **Preventing Bullying and Harassment:** Hate speech often leads to bullying and harassment, both online and offline. Empowering adults to address it helps create safer spaces, reducing instances of bullying.
- **Upholding Human Rights:** Hate speech violates basic human rights like dignity and equality. Equipping adults to confront it supports global efforts to uphold human rights and promote social justice.
- **Preparing Children for a Diverse World:** In today's interconnected world, children need to navigate diverse perspectives respectfully. Adults trained to address hate speech can help prepare children to engage thoughtfully with different cultures.
- **Fostering Responsible Digital Citizenship:** As hate speech spreads online, it threatens children's digital safety. Teaching adults to promote responsible digital citizenship helps create safer digital environments where hate speech is less tolerated.

Preventing hate speech among children and young people requires a multifaceted approach that addresses its root causes, fosters critical thinking and empathy, and empowers active citizenship. By investing in education, awareness-raising and inclusive policies, we can build a future where all children and young people can thrive in environments free from discrimination and hatred.

The following section presents a number of suggestions for methods and analogies that can be employed in the field of education in particular.

Ethnographic analogies to engage youth in understanding hate speech

Trash can analogy

Drawing from an ethnographic example, we can imagine a cartoon where a group of young children carelessly litter instead of using a trash can. Each piece of trash they discard may seem insignificant on its own, but as more accumulates, it creates a larger problem, polluting their surroundings. Similarly, individuals who casually use hate speech – perhaps thinking that a single insult or stereotype is harmless – contribute to a broader atmosphere of harm. Over time, these words accumulate, negatively affecting those around them, often in ways that go unnoticed at first.

Balloon analogy

Inspired by ethnographic examples, we can imagine a cartoon where young people inflate balloons, enjoying the process without realizing that when a balloon bursts, it startles and disturbs those around them. At first, inflating the balloons seems harmless, even entertaining, but as more balloons pop, the noise creates discomfort and disruption. Similarly, on social media, people often engage in discussions within their own **"bubbles"**, reinforcing shared beliefs without outside perspectives. These echo chambers are not accidental – they are actively reinforced by platform algorithms that prioritize engagement, amplifying content that triggers strong emotional reactions, including outrage and hostility. As a result, individuals are repeatedly exposed to the same perspectives, inflating their ideological "bubbles" without challenge. Additionally, trolls and bots intensify these dynamics by artificially boosting divisive content, further solidifying these self-contained spheres. The more these narratives circulate unchecked, the more extreme they become – like overinflated balloons on the verge of bursting. When these bubbles finally break into the wider public sphere, the harmful rhetoric spills over, affecting individuals and communities beyond the original group, often with real-world consequences.

The ripple effect analogy

Imagine a stone thrown into a calm lake. The impact creates ripples that expand outwards, touching everything in their path. A single hateful remark or discriminatory stereotype, much like a stone thrown into a lake, creates ripples that extend far beyond the initial moment. Hate speech does not **remain isolated** – it influences individuals, affects communities, and ultimately shapes society's structure. When discriminatory narratives are reinforced through history books and educational curricula, they legitimize exclusion and bias, making them even harder to dismantle.

Educational tools for countering hateful narratives

Picture books and stories

Children's books and stories are powerful tools for challenging hate speech, dismantling harmful historical narratives and preventing the reinforcement of prejudices through education. By introducing diverse perspectives, addressing discriminatory stereotypes and fostering critical thinking, stories can help young readers develop empathy and resilience against harmful discourse. When hate speech and exclusionary narratives are embedded in education, they shape children's understanding of history, identity and others around them. By intentionally curating inclusive and critical storytelling, education can become a tool for empowerment rather than reinforcing discrimination.

Gamification: Learning to identify and counter hate speech

Games are a valuable educational tool that can be used to raise awareness among children and young people about hate speech and discrimination, while also reducing their biases toward different identities. By integrating elements such as challenges, rewards and role-playing scenarios, gamification enables participants to engage with real-world issues in a safe and immersive environment. Games can be designed to help youth recognize how misinformation, biased discourse and historical prejudices shape public opinion.

A conceptual example of how gaming could be used to raise awareness about discriminatory narratives is the imagined game *Logic Defender: The Quest Against Hate Speech!* In this game, players would navigate various online environments – such as social media platforms, forums and comment sections – where they encounter hate speech disguised as logical arguments. By recognizing fallacies, players would learn how misinformation spreads and how historical biases are reinforced through discourse.

Hackathons – workshops for digital activism

Hackathons and digital workshops allow young people to explore innovative ways to combat online hate speech and misinformation. By developing strategies to counter discriminatory narratives and designing digital tools that identify and disrupt harmful discourse, participants can take an active role in reshaping digital spaces into more inclusive environments.

Creative resistance: Art as a tool to empower youth against hateful narratives

Hate speech and discriminatory narratives deeply impact children and youth, influencing how they see themselves, others, and historical and social discourses. Art can serve as a shield against these adverse effects, providing a means of expression and resistance, and redefining the narratives that shape our surroundings. By engaging young people in creative methods to challenge discrimination, they can express their perspectives, contribute to more inclusive narratives, and foster a culture that resists exclusion.

Painting: Reimagining narratives

Through painting, young people can express the emotional outcomes of hate speech, visualize the destruction of discrimination, and illustrate a future based on values such as equality and diversity. Art workshops may allow them to challenge dominant narratives and imagine new ones that reflect diverse histories and perspectives.

Theater: Challenging narratives

Theater provides a creative space for young people to explore different perspectives, express themselves freely, and critically engage with social issues, including discrimination and hate speech. By immersing themselves in diverse roles and narratives, they gain a deeper understanding of how words and actions affect individuals and communities.

Participating in theater enhances communication skills, builds self-confidence and encourages teamwork, empowering youth to articulate their thoughts and navigate complex social dynamics. Rather than placing the burden on them to challenge discrimination directly, theater helps them develop the tools to recognize bias, question harmful narratives and advocate for inclusivity in ways that feel natural to them. Through storytelling and performance, they can experiment with self-expression, foster empathy, and appreciate the value of dialogue and diverse voices in shaping a more understanding discourse.

Digital storytelling: Navigating and reshaping narratives

Young people increasingly engage with the world through digital platforms – whether for news, social interaction or entertainment – it is essential to equip them with the tools to tell their own stories effectively, while critically analyzing the narratives they encounter. Digital storytelling provides an opportunity to express perspectives, challenge discrimination and counter harmful narratives, all while developing media literacy skills that are crucial in today's information landscape.

Through digital storytelling, young people can engage with discussions on the impact of hate speech, analyze how biased narratives are constructed, and experiment with presenting more inclusive perspectives. By exploring different storytelling techniques, they gain the tools to express themselves while critically examining the messages they encounter in digital spaces.

At the same time, digital storytelling equips youth with the ability to recognize disinformation and misinformation campaigns that manipulate similar techniques to spread false or harmful narratives. By analyzing how media shapes public perception, young people can become more discerning consumers and producers of content. Workshops in digital storytelling provide a space for both creative expression and critical engagement, ensuring that youth are not only participants in the digital world but also active shapers of the narratives that define it.

An exemplary handbook from Turkey, developed with the objective of enhancing the critical digital literacy of children and young people, provides an excellent illustration of how digital literacy can be incorporated into education to empower students with the ability to critically assess digital content, recognize misinformation, and develop responsible online engagement.

Critical Digital Literacy in Education: A Handbook by and for Teachers[62] was prepared in collaboration with **Teachers Network**[63] and **Teyit.org**[64], with contributions by 39 teachers of various subjects from 19 different cities of Turkey. The book addresses the impact of critical digital literacy skills on educational environments and establishes a connection with confirmationism. It includes chapters that will nourish educators' awareness and interest in the field, as well as suggestions for activities that can be applied in learning environments.

[62] **Öğretmen Ağı.** (2023). *Edokitap: Educational resources for teachers*. Öğretmen Ağı. https://www.ogretmenagi.org/sites/www.ogretmenagi.org/files/publications/edokitap_eng_2_03_2023.pdf

[63] **Teachers' Network.** (2024). *Teachers' Network – A collaborative learning platform for teachers.* https://www.ogretmenagi.org/en

[64] **Teyit.** (2024). *Teyit – Fact-checking and verification platform.* https://en.teyit.org/

The Hrant Dink Foundation's ASULIS Discourse, Dialogue, and Democracy Laboratory has developed an **Inclusive Discourse Workshop**[65] to address the impact of language on social inclusion and equity. The workshop is based on the Foundation's **Media Watch on Hate Speech**[66] project, which has been ongoing since 2009. The workshop aims to equip participants with the knowledge and tools needed to recognize and challenge discriminatory language patterns in everyday communication. Combining theoretical insights with practical exercises, the workshop fosters a deeper understanding of how language can perpetuate exclusion or promote inclusivity. The workshop's content is regularly updated by tracking linguistic changes, as well as incorporating insights from discourse studies and related reports. By encouraging participants to adopt more inclusive communication practices, the workshop contributes to creating fairer and more equitable environments in both personal and professional contexts, reflecting the Foundation's long-standing commitment to fostering dialogue and understanding.

[65] **Hrant Dink Foundation.** (2022). *Inclusive discourse workshop*. Hrant Dink Foundation. https://hrantdink.org/en/asulis/announcements/4377-inclusive-discourse-workshop

[66] **Hrant Dink Foundation.** (2016). *Media watch on hate speech*. Hrant Dink Foundation. https://hrantdink.org/en/asulis/activities/projects/media-watch-on-hate-speech

In lieu of conclusion

This report captures the collective insights and interdisciplinary discussions that emerged from a year-long collaboration among experts, dedicated to understanding and addressing hate speech and discriminatory discourse. By examining the identification and categorization of hate speech, the challenges of detection, and strategies for countering it beyond policy measures, this publication provides a comprehensive overview of the complexities involved in tackling this issue.

The contributions in this report highlight the need for nuanced definitions, ethical approaches and multi-layered interventions that go beyond legal and regulatory frameworks. The discussions emphasize the importance of interdisciplinary collaboration, civil society engagement and educational initiatives in fostering inclusive and constructive discourse.

We hope that this publication serves as a valuable resource for researchers, activists, policymakers, and all those working to promote social cohesion, equality, and dialogue. The insights and recommendations presented here not only reflect the work of this network but also contribute to ongoing efforts toward a more inclusive public discourse.

Alex Mahadevan

Alex Mahadevan is the director of MediaWise at Poynter, leading AI initiatives and misinformation research. He has trained thousands of students, older adults, and journalists in online verification and media literacy. He co-wrote Poynter's AI ethics guide, led its first Summit on AI, Ethics, and Journalism, and co-leads Stanford's Empowering Diverse Digital Citizens Lab. Mahadevan has conducted workshops on generative AI, OSINT, and emerging technologies worldwide.

Arzucan Özgür

Arzucan Özgür is a faculty member in Computer Engineering at Boğaziçi University and co-director of the TABI Lab. Her research focuses on bioinformatics and natural language processing, developing algorithms for human and biological languages. She holds a Ph.D. from the University of Michigan and M.S. and B.S. degrees from Boğaziçi University. Previously, she worked at Istanbul Technical University and is a member of AILAB.

Ayşecan Terzioğlu

Ayşecan Terzioğlu is a faculty member at Sabancı University's Cultural Studies and Gender Studies Programs. She earned her Ph.D. in Anthropology from the City University of New York after completing her B.A. and M.A. at Boğaziçi University. Her research focuses on the anthropology of the Middle East, biopolitics, health inequalities, gender, and social studies on temporality and spatiality. She has published extensively and serves on the boards of RHWG, AMEA, and JOTSA. Her research focuses on the anthropology of the Middle East, biopolitics, health inequalities, gender, social studies on temporality and spatiality and hate speech.

Berrin Yanıkoğlu

Berrin Yanıkoğlu is a Professor of Computer Science and founding Director of the Center of Excellence in Data Analytics (VERIM) at Sabancı University. She earned a double major from Boğaziçi University and a Ph.D. from Dartmouth College. Before joining Sabancı University in 2000, she worked at Rockefeller University, Xerox Imaging Systems, and IBM Almaden. Her research focuses on computer vision and machine learning, including hate speech detection, image understanding, handwriting recognition, and text summarization.

Claudia von Vacano

Dr. Claudia von Vacano is an expert in algorithmic fairness, transparency, and explainability, focusing on hate speech. As Executive Director of UC Berkeley's D-Lab and related programs, she supports over 6,000 scholars annually through data-intensive research and training. She co-leads a multi-campus NSF grant on educational and public health equity and developed UC Berkeley's Digital Humanities Summer Minor. She also leads the Online Hate Index with the Anti-Defamation League and holds a Ph.D. from UC Berkeley and a Master's from Stanford University.

Didar Akar

Didar Akar is an Associate Professor of Linguistics at Boğaziçi University. She holds a Ph.D. in Linguistics from the University of Michigan. Her teaching and research areas include discourse analysis, conversation analysis, sociolinguistics and pragmatics. She focuses on language and gender, language and identity, and most recently, social media and hate speech.

Eser Selen

Eser Selen is an Associate Professor of Communication Design at Özyeğin University, specializing in feminisms, performance studies, queer theory, and contemporary art. Her research has been published in journals such as *Gender, Place & Culture* and *Women & Performance.* She is finalizing her first monograph, *Contesting Gender and Sexuality through Performance*, set for publication in 2026 by the University of Edinburgh Press. A multimedia artist, her work in performance, installation, and video has been exhibited internationally.

Houda Bouamor

Dr. Houda Bouamor is an Assistant Teaching Professor of Information Systems at Carnegie Mellon University in Qatar, and a Research Scientist working in the fields of natural language processing and machine learning. Dr. Bouamor is an expert in natural language processing and computational linguistics. She received her Ph.D. in Computer Science from Paris-Sud University. Dr. Bouamor is an active member of the research community serving as program committee member in over 20 conferences. Her main research interest revolves around statistical machine translation.

İlayda Ece Ova

İlayda Ece Ova is a researcher and program analyst specializing in gender equality, civic engagement, and digital activism.

Metin V. Bayrak

Metin V. Bayrak studied Philosophy at Hacettepe University and has focused on applied philosophy since 2008. Through Opus Noesis, which he founded in 2014, he promotes applied philosophy in Turkey and teaches at Galatasaray University and Istanbul Okan University. He leads certificate programs in applied philosophy and P4C. In 2021, he founded Opus Kitap to publish works on applied philosophy.

Onur Varol

Onur Varol is an Assistant Professor in the Computer Science Department at Sabancı University and affiliated with the Center of Excellence in Data Analytics (VERIM) . He leads the VIRAL Lab, focusing on computational social science, networks, and machine learning. His research includes algorithmic fairness, social bot detection, and modeling user interactions across online platforms.

Peter Adams

Peter Adams is The News Literacy Project’s Senior Vice President of Research and Design, and has been with the organization since 2009. He began his career in education as a classroom teacher in the New York City schools. He has also worked as a trainer with the New York City Teaching Fellows Program, a youth media after-school instructor in the Chicago public schools and an adjunct instructor at Roosevelt University and Chicago City Colleges.

Roser Morante

Dr. Roser Morante is a researcher at UNED University, affiliated with the Natural Language Processing and Information Retrieval group. She holds a Ph.D. in Computational Linguistics from Tilburg University and has published over 80 research papers. Her academic work includes teaching, organizing workshops, and editing journal issues, such as the *Computational Linguistics* Special Issue on Modality and Negation. She has co-organized events on sexism identification in social networks and her research focuses on semantic and discursive aspects of meaning.

Susan Benesch

Susan Benesch founded and directs the Dangerous Speech Project, which studies speech that can incite violence and explores ways to prevent harm without limiting free expression. She co-founded the Coalition for Independent Technology Research to support related work and advises tech companies on content governance. A trained human rights lawyer from Yale University, Benesch teaches at American University and is a Faculty Associate at Harvard’s Berkman Klein Center. Her current research focuses on grassroots responses to online hatred and design friction.

Tirşe Erbaysal Filibeli

Associate Professor Tirşe Erbaysal Filibeli holds MA and Ph.D. degrees in Media and Communication Studies from Galatasaray University. She has been a researcher and coordinator for the Turkey team of the Media Pluralism Monitoring Project since 2016 and worked on the Ermiscom Project from 2020-2023. She has co-edited books on topics such as fake news, digital media literacy, and journalism in conflict, and her research covers areas such as hate speech, populism, and digital capitalism. Filibeli has been leading the New Media Department at Bahçeşehir University Faculty of Communication since 2018.

Tommaso Caselli

Dr. Tommaso Caselli is an Assistant Professor in Computational Semantics at the University of Groningen's Center for Language and Cognition. His research focuses on event extraction, framing, hate speech detection, and misinformation countering. He co-edited *Computational Analysis of Storylines* (CUP, 2021) and has organized several NLP semantic evaluation campaigns for English and Italian. Dr. Caselli has received multiple awards for his papers and is the coordinator of the AI and Language theme at the Jantina Tammes School of Digital Society.

Yasemin İnceoğlu

Yasemin İnceoğlu graduated in English Language and Literature from Istanbul University and received her MA and Ph.D. in Journalism at Marmara University, where she became a professor in 1999. She served as the Head of the Journalism Department and Dean of the Faculty of Communication at Galatasaray University until her retirement in 2016. İnceoğlu has been involved with international organizations such as UNESCO, ILAD, and the Council of Europe, and has lectured at institutions like EUI and UCL. She is the author of several books on media, minorities, hate speech, and populism, and currently works in the Department of Media and Communication at the London School of Economics.

Zeynep Özarslan

Zeynep Özarslan, holding a Ph.D. in Communication Sciences, is a Professor of Communication Studies at Çukurova University. She has taught courses in film, communication, and new media studies since 1998, supervising graduate theses. Her research interests include new media, film studies, the sociology of communication, and the creative industries, with recent publications focusing on digital games, celebrity studies, digital violence, and hate speech. Outside of academia, she collaborates with CSOs to combat hate speech, digital violence, and promote internet rights.

References

Acker, J. (2012). Gendered organizations and intersectionality: Problems and possibilities. *Equality, Diversity and Inclusion: An International Journal*, 31(3), 214–224. https://doi.org/10.1108/02610151211209072

Alessi, E. J., Kahn, S., & Van Der Horn, R. (2017). A Qualitative Exploration of the Premigration Victimization Experiences of Sexual and Gender Minority Refugees and Asylees in the United States and Canada. *Journal of Sex Research*, 54(7), 936–948. https://doi.org/10.1080/00224499.2016.1229738

Arın, I., Işık, Z., Kutal, S., Dehghan, S., Özgür, A., & Yanikoğlu, B. (2023, July). SIU2023-NST-hate speech detection contest. In *2023 31st Signal Processing and Communications Applications Conference (SIU)* (pp. 1–4). IEEE.

Aroyo, L., & Welty, C. (2015). Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1), 15–24. https://doi.org/10.1609/aimag.v36i1.2564

Artstein, R., & Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4), 555–596. https://doi.org/10.1162/coli.07-034-R2

Awal, M. R., Cao, R., Lee, R. K. W., & Mitrović, S. (2020). On analyzing annotation consistency in online abusive behavior datasets. *arXiv preprint arXiv:2006.13507*. https://doi.org/10.48550/arXiv.2006.13507

Badali, J. J. (2019). Migrants in the closet: LGBT migrants, homonationalism, and the right to refuge in Serbia. *Journal of Gay & Lesbian Social Services*, 31(1), 89–119. https://doi.org/10.1080/10538720.2019.1548330

Balčytienė, A., & Juraitė, K. (2022). Monitoring media pluralism in the digital era: Application of the Media Pluralism Monitor in Lithuania in the year 2021. *European University Institute, Centre for Media Pluralism and Media Freedom.* https://doi.org/10.2924/EJLS.2022.015

Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Rangel Pardo, F. M., Rosso, P., & Sanguinetti, M. (2019). SemEval-2019 Task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation* (pp. 54–63). https://doi.org/10.18653/v1/S19-2007

Bleyer-Simon, K., Brogi, E., Carlini, R., Da Costa Leite Borges, D., Kermer, J. E., Nenadic, I., Palmer, M., Parcu, P. L., Reviglio Della Venaria, U., Trevisan, M., Verza, S., & Žuffová, M. (2024). Monitoring media pluralism in the digital era: Application of the media pluralism monitor in the European member states and in candidate countries in 2023. *EUI, RSC, Research Project Report, Centre for Media Pluralism and Media Freedom (CMPF).*
https://hdl.handle.net/1814/77028

Bosco, C., Felice, D. O., Poletto, F., Sanguinetti, M., & Maurizio, T. (2018). Overview of the EVALITA 2018 hate speech detection task. In *CEUR Workshop Proceedings* (Vol. 2263, pp. 1–9).

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Sorelle & Wilson (Eds.), *Proceedings of the 2018 Conference* (pp. 1–15).

Cabitza, F., Campagner, A., & Basile, V. (2023, June). Toward a perspectivist turn in ground truthing for predictive computing. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 37, No. 6, pp. 6860–6868).

Chun, J. J., Lipsitz, G., & Shin, Y. (2013). Intersectionality as a social movement strategy: Asian immigrant women advocates. *Signs: Journal of Women in Culture and Society, 38*(4), 917–940. http://www.jstor.org/stable/10.1086/669575

Council of Europe. (1997). Recommendation No. R (97) 20 of the Committee of Ministers to Member States on "Hate Speech." In *Recommendations and Declarations of the Committee of Ministers in the Field of Media and Information Society* (pp. 106–108). *Strasbourg: Council of Europe.*

Council of Europe. (2016, December). Seminar on combating sexist hate speech: Report (10-12 February 2016, European Youth Centre, Strasbourg). *Council of Europe.* https://rm.coe.int/16806cc316

Crenshaw, K. (1989). Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *University of Chicago Legal Forum,* 139–167. https://chicagounbound.uchicago.edu/cgi/viewcontent.cgi?article=1052&context=uclf

Crenshaw, K. (2005). Mapping the margins: Intersectionality, identity politics, and violence against women of color. In R. K. Bergen, J. L. Edleson, & C. M. Renzetti (Eds.), *Violence against women: Classic papers* (pp. 282–313). Pearson Education New Zealand.

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media* (pp. 512–515). Association for the Advancement of Artificial Intelligence. https://doi.org/10.1609/icwsm.v11i1.14955

Del Vigna, F., Cimino, A., Dell'Orletta, F., Petrocchi, M., & Tesconi, M. (2017). Hate me, hate me not: Hate speech detection on Facebook. In *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17),* Venice, Italy, 86–95.

Djuric, N., Zhou, J., Morris, R., Grbovic, M., Radosavljevic, V., & Bhamidipati, N. (2015). Hate speech detection with comment embeddings. *WWW'15 Companion: Proceedings of the 24th International Conference on World Wide Web.* 29–30. https://doi.org/10.1145/2740908.2742760

Eduardo Sánchez, M., & Eduardo, M. (2013). Latino Lesbian, Gay, Bisexual, and Transgender Immigrants in the United States. *Journal of LGBT Issues in Counseling*. https://doi.org/10.1080/15538605.2013.785467

Edward, J. A., Kershaw, S., Grimm, B., Wehner, L., & Murphy, D. (2020). A qualitative exploration of the integration experiences of LGBTQ refugees who fled from the Middle East, North Africa, and Central and South Asia to Austria and the Netherlands. *Sexuality Research and Social Policy*. https://doi.org/10.1007/s13178-018-0364-7

Erbaysal Filibeli, T., & Ertuna, C. (2021). Sarcasm beyond hate speech: Facebook comments on Syrian refugees in Turkey. *International Journal of Communication, 15,* 2236–2259. https://ijoc.org/index.php/ijoc/article/view/16582

European Commission. (2020). The European Union's Strategy and Action Plan on Hate Speech. https://ec.europa.eu/info/policies/justice-and-fundamental-rights/anti-discrimination/hate-speech_en

Faloppa, F., Gambacorta, A., Odekerken, R., & van der Noordaa, R. (2023). Study on preventing and combating hate speech in times of crisis. Council of Europe.
https://rm.coe.int/-study-on-preventing-and-combating-hate-speech-in-times-of-crisis/1680ad393b

Fassinger, R. E., & Arseneau, J. R. (2007). "I'd rather get wet than be under that umbrella": Differentiating the experiences and identities of lesbian, gay, bisexual, and transgender people. In K. J. Bieschke, R. M. Perez, & K. A. DeBord (Eds.), *Handbook of counseling and psychotherapy with lesbian, gay, bisexual, and transgender clients* (2nd ed., pp. 19–49). American Psychological Association. https://doi.org/10.1037/11482-001

Feinberg, J. (1983). Obscene words and the law. *Law and Philosophy, 2*(2), 139–161.

Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR),* 51(4), 1–30. https://doi.org/10.1145/3236009.3236010
Fortuna, P., Rocha da Silva, J., Wanner, L., & Nunes, S. (2019). A hierarchically-labeled Portuguese hate speech dataset. In *Proceedings of the Third Workshop on Abusive Language Online* (pp. 94–104). https://doi.org./10.18653/v1/W19-3510

Fortuna, P., Soler, J., & Wanner, L. (2020, May). Toxic, hateful, offensive or abusive? What are we really classifying? An empirical analysis of hate speech datasets. In *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 6786–6794). European Language Resources Association. https://aclanthology.org/2020.lrec-1.838.pdf

*Gambäck, B., & S*ikdar, U. K. (2017). Using convolutional neural networks to classify hate-speech. *In Proceedings of the First Workshop on Abusive Language Online (ALW1),* 85–90. https://doi.org./10.18653/v1/W17-3013

Gao, S. (S.), Brandt, S. A., & Stults, C. B. (2023). Internalized transphobia and self-concept clarity among transgender and gender-nonconforming young adults: Characteristics, associations, and the mediating role of self-esteem. *Psychology of Sexual Orientation and Gender Diversity.* Advance online publication.
https://doi.org/10.1037/sgd0000691

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumée, III, H., & Crawford, K. (2018). Datasheets for datasets. *arXiv,* 1–17. https://arxiv.org/abs/1803.09010

Gelber, K. (2021). Differentiating hate speech: A systemic discrimination approach. *Critical Review of International Social and Political Philosophy,* 24(4), 393–414. https://doi.org/10.1080/13698230.2019.1576006

Gibert, O., Perez, N., García-Pablos, A., & Cuadros, M. (2018). Hate speech dataset from a white supremacy forum. In D. Fišer, R. Huang, V. Prabhakaran, R. Voigt, Z. Waseem, & J. Wernimont (Eds.), *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)* (pp. 11–20). Association for Computational Linguistics. https://aclanthology.org/W18-5102/

Gitari, N. D., Zuping, Z., Damien, H., & Long, J. (2015). A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10(4), 215–230. https://doi.org/10.14257/ijmue.2015.10.4.21

GLAAD. (2024). Promoting Positive Narratives and Monitoring Anti-LGBTQ+ Media Portrayals. https://www.glaad.org/

Gowin, M., Taylor, E. L., Dunnington, J., Alshuwaiyer, G., & Cheney, M. K. (2017). Needs of a silent minority: Mexican transgender asylum seekers. *Health Promotion Practice, 18*(3), 332–340. https://doi.org/10.1177/1524839917692750

Hagendorff, T. (2020). Minds and Machines, 30, 99–120. https://doi.org/10.1007/s11023-020-09517-8

Holznagel, B., & Kalbhenn, J. C. (2023). Monitoring media pluralism in the digital era: Application of the media pluralism monitor in the European Union, Albania, Montenegro, the Republic of North Macedonia, Serbia, and Turkey in the year 2022: Country report: Germany. European University Institute. https://hdl.handle.net/1814/75723

Hopkinson, R. A., Keatley, E., Glaeser, E., Erickson-Schroth, L., Fattal, O., & Nicholson Sullivan, M. (2017). Persecution experiences and mental health of LGBT asylum seekers. *Journal of Homosexuality, 64*(12), 1650–1666.
https://doi.org/10.1080/00918369.2016.1253392

Hosseinmardi, H., Mattson, S. A., Rafiq, R. I., Han, R., Lv, Q., & Mishra, S. (2015). Analyzing labeled cyberbullying incidents on the Instagram social network. *International Conference on Social Informatics*, 49–66. https://doi.org/10.1007/978-3-319-27433-1_4

İnanç Özgürlüğü Girişimi. (2021). *Rabat eylem planı (Turkish)*.
https://inancozgurlugugirisimi.org/wp-content/uploads/2021/02/Rabat-Eylem-Plani-Turkce.pdf

İnceoğlu, Y., Erbaysal Filibeli, T., Ertuna, C., & Çenberli, Y. (2022). Monitoring media pluralism in the digital era: Application of the Media Pluralism Monitor in Turkey in the year 2022. European University Institute, Centre for Media Pluralism and Media Freedom. https://hdl.handle.net/1814/75744

Jahan, M. S., & Oussalah, M. (2023). A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, 546, 126232. https://doi.org/10.1016/j.neucom.2023.126232

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1, 389–399. https://doi.org/10.1038/s42256-019-0088-2

Kaya, A. (2023). The world's leading refugee host, Turkey, has a complex migration history. *Migration Policy Institute*. https://www.migrationpolicy.org/article/turkey-migration-history

Kılıç, O. (2023). Queer cybercultures and digital activism: Transformations in LGBTQ+ digital spaces. *Lambda Nordica, 28*(2-3), 902. https://doi.org/10.34041/ln.v28.902

Kindermann, D. (2023). Against 'Hate Speech'. *Journal of Applied Philosophy*, 40(5). https://doi.org/10.1111/japp.12648

Kocon, J., Figas, A., Gruza, M., Pulchaska, D., Kajdanowicz, T., & Kazienko, P. (2021). Offensive, aggressive, and hate speech analysis: From data-centric to human-centered approach. *Information Processing and Management*, 58, 1-26. https://doi.org/10.1016/j.ipm.2021.102643

Kwok, I., & Wang, Y. (2013). Locate the hate: Detecting tweets against blacks. *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence (AAAI-13)*, 1621–1622. https://doi.org/10.1609/aaai.v27i1.8539

Leader Maynard, S., & Benesch, S. (2018). Dangerous speech and hate speech: Definitions, characteristics, and analysis. *International Journal of Communication*, 12, 2740–2759. https://ijoc.org/index.php/ijoc/article/view/8200

Leonardelli, E., Basile, V., & Bosco, C. (2021). Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement. *Proceedings of the 11th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 167–176. https://aclanthology.org/2021.wassa-1.17

Liebe, U., & Beyer, H. (2021). Examining discrimination in everyday life: A stated choice experiment on racism in the sharing economy. *Journal of Ethnic and Migration Studies*, 47(9), 2065–2088. https://doi.org/10.1080/1369183X.2019.1710118

Loebbecke, C., Luong, A. C., & Obeng-Antwi, A. (2021). AI for tackling hate speech. ECIS 2021 Research-in-Progress Papers, 10.

Madukwe, K., Gao, X., & Xue, B. (2020, November). In data we trust: A critical analysis of hate speech detection datasets. In *Proceedings of the Fourth Workshop on Online Abuse and Harms* (pp. 150–161). Association for Computational Linguistics.

Manzi, F., & Heilman, M. E. (2021). Breaking the glass ceiling: For one and all? *Journal of Personality and Social Psychology*, 120(2), 257–277. https://doi.org/10.1037/pspa0000260

McClain, M., & Waite-Wright, O. (2015). The LGBT community in Turkey: Discrimination, violence, and the struggle for equality. *Creighton International and Comparative Law Journal*, *7*(1), 152–176.

McLean, L. L. (2021). Internalized homophobia and transphobia. In E. M. Lund, C. Burgess, & A. J. Johnson (Eds.), *Violence against LGBTQ+ persons*. Springer. https://doi.org/10.1007/978-3-030-52612-2_3

Mediaweek. (2022, 26 October). *TikTok and AAP partner to empower content creators to combat misinformation.* Mediaweek. https://www.mediaweek.com.au/tik-tok-and-aap-partner-to-empower-creators-to-fight-misinformation/

Mehta, H., & Passi, K. (2022). Social media hate speech detection using explainable artificial intelligence (XAI). *Algorithms*, 15(8), 291. https://doi.org/10.3390/a15080291

Morales, E. (2013). Latino lesbian, gay, bisexual, and transgender immigrants in the United States. *Journal of LGBT Issues in Counseling*, *7*(2), 172–184. https://doi.org/10.1080/15538605.2013.785467

Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2020). From what to how: An initial review of publicly available AI ethics tools, methods, and research to translate principles into practices. *Science and Engineering Ethics*, *26*(4), 2141–2168. https://doi.org/10.1007/s11948-019-00165-5

Morstatter, F., Pfeffer, J., Liu, H., & Carley, K. (2013). Is the sample good enough? Comparing data from Twitter's streaming API with Twitter's Firehose. *Proceedings of the 7th International AAAI Conference on Web and Social Media (ICWSM 2013)* (pp. 400–408). Association for the Advancement of Artificial Intelligence (AAAI). https://doi.org/10.1609/icwsm.v7i1.14401

Musto, C., Semeraro, G., Polignano, M., & Stranisci, M. (2016). Modeling community behavior through semantic analysis of social data: The Italian Hate Map experience. In *Proceedings of the 8th International Conference on Knowledge Discovery and Information Retrieval (KDIR 2016)* (pp. 343–350). SCITEPRESS – Science and Technology Publications.

Nadal, K. (2008). Preventing racial, ethnic, gender, sexual minority, disability, and religious microaggressions: Recommendations for promoting positive mental health. *Prevention in Counseling Psychology: Theory, Research, Practice and Training, 2*(1), 22–27.

Nadal, K. L., Issa, M.-A., Leon, J., Meterko, V., Wideman, M., & Wong, Y. (2011a). Sexual orientation microaggressions: "Death by a thousand cuts" for lesbian, gay, and bisexual youth. *Journal of LGBT Youth, 8*(3), 234–259. https://doi.org/10.1080/19361653.2011.584204

Nadal, K. L., Wong, Y., Issa, M.-A., Meterko, V., Leon, J., & Wideman, M. (2011b). Sexual orientation microaggressions: Processes and coping mechanisms for lesbian, gay, and bisexual individuals. *Journal of LGBT Issues in Counseling,* 5(1-2), 21–46. https://doi.org/10.1080/15538605.2011.554606

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., & Chang, Y. (2016). Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web (WWW 2016)* (pp. 145–153). ACM

Nockleby, J. T. (2000). Hate speech. *Encyclopedia of the American Constitution, 3,* 1277–1279.

Novak, P. K., Scantamburlo, T., Pelican, A., Cinelli, M., Mozetič, I., & Zollo, F. (2022). Handling disagreement in hate speech modelling. *Communications in Computer and Information Science, 1602,* 681–695. https://doi.org/10.1007/978-3-031-08974-9_54

O'Driscoll, J. (2020). Offensive language: Taboo, offence and social control. Bloomsbury Publishing.

Office of the United Nations High Commissioner for Human Rights (OHCHR). (2012). Rabat Plan of Action on the prohibition of advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence. OHCHR. https://www.ohchr.org/sites/default/files/Rabat_draft_outcome.pdf

Özdüzen, A., & Korkut, U. (2020). 'Refugees are not welcome': Digital racism, online place-making and the evolving categorization of Syrians in Turkey. *Environment and Planning C: Politics and Space, 39*(5), 851–869. https://journals.sagepub.com/doi/10.1177/1461444820956341

Pasquetto, I. V., Priedhorsky, R., Lazer, D., & Terveen, L. (2020). Ethical and transparency challenges in data collection for misinformation research. *Journal of Information Ethics, 29*(1), 58–77. https://doi.org/10.1016/j.jinfo.2020.01.004

Pfeffer, J., Zorbach, T., & Carley, K. M. (2023). Challenges in collecting data on social media platforms: A case study of X/Twitter. *Proceedings of the 2023 International Conference on Social Media & Society* (pp. 1–10).

Piwowarczyk, L., Fernandez, P., & Sharma, A. (2017). Seeking asylum: Challenges faced by the LGB community. *Journal of Immigrant and Minority Health,* 19(4), 723–732. https://doi.org/10.1007/s10903-016-0363-9

Plaza, L., et al.,(2023). Overview of EXIST 2023: sEXism identification in social networks. In J. Kamps et al.,(Eds.), *Advances in information retrieval: ECIR 2023* (pp. 1–15). Springer. https://doi.org/10.1007/978-3-031-28241-6_68

Plank, B., Hovy, D., & Søgaard, A. (2014). Learning part-of-speech taggers with inter-annotator agreement loss. *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics,* 742–751. https://aclanthology.org/E14-1078

Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., & Patti, V. (2021). Resources and benchmark corpora for hate speech detection: A systematic review. *Language Resources and Evaluation,* 55, 477–523. https://doi.org/10.1007/s10579-020-09502-8

Poynter Institute. (2020, December 10). *Poynter's MediaWise training significantly increases people's ability to detect disinformation, new Stanford study finds.* Poynter Institute. https://www.poynter.org/news-release/2020/poynters-mediawise-training-significantly-increases-peoples-ability-to-detect-disinformation-new-stanford-study-finds/

Rebillard, F., & Sklower, J. (2022). Monitoring media pluralism in the digital era: Application of the Media Pluralism Monitor in France in the year 2021. *European University Institute, Centre for Media Pluralism and Media Freedom.*
https://hdl.handle.net/1814/74750

Robles Carrillo, M. (2020). Artificial intelligence: From ethics to law. *Telecommunications Policy*. https://doi.org/10.1016/j.telpol.2020.101937

Rodríguez-Sánchez, F., Carrillo-de-Albornoz, J., Plaza, L., Gonzalo, J., Rosso, P., Comet, M., & Donoso, T. (2021). Overview of EXIST 2021: sEXism identification in social networks. *Procesamiento del Lenguaje Natural,* 67, 195–207.

Rodríguez-Sánchez, F., Carrillo-de-Albornoz, J., Plaza, L., Mendieta-Aragón, A., Marco-Remón, G., Makeienko, M., Plaza, M., Gonzalo, J., Spina, D., & Rosso, P. (2022). Overview of EXIST 2022: sEXism identification in social networks. *Procesamiento del Lenguaje Natural,* 69, 229–240.

Royakkers, L., Timmer, J., Kool, L., & van Est, R. (2018). Societal and ethical issues of digitization. *Ethics and Information Technology, 20*(2), 127–142. https://doi.org/10.1007/s10676-018-9452-x

Röttger, P., Vidgen, B., Nguyen, D., Waseem, Z., Margetts, H., & Pierrehumbert, J. (2021). HateCheck: Functional tests for hate speech detection models. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics* (pp. 41–58). Association for Computational Linguistics.

Röttger, P., Vidgen, B., Hovy, D., & Pierrehumbert, J. (2022). Two contrasting data annotation paradigms for subjective NLP tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.* Association for Computational Linguistics.

Ruitenbeek, W., Zwart, V., van der Noord, R., Gnezdilov, Z., & Caselli, T. (2022). "Zo Grof!": A Comprehensive Corpus for Offensive and Abusive Language in Dutch. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)* (pp. 40–56). Association for Computational Linguistics. https://aclanthology.org/2022.woah-1.5/

Russell, E. L. (2020). Hate in language, hate and language. In K. Hall & R. Barrett (Eds.), *The Oxford handbook of language and sexuality*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190212926.013.67

Sachdeva, P. S., Barreto, R., von Vacano, C., & Kennedy, C. J. (2022, June). Assessing annotator identity sensitivity via item response theory: A case study in a hate speech corpus. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1585–1603). Association for Computing Machinery.

Salminen, J., Almerekhi, H., Milenkovič, M., Jung, S. G., An, J., Kwak, H., & Jansen, B. J. (2018). Anatomy of online hate: Developing a taxonomy and machine learning models for identifying and classifying hate in online news media. *Proceedings of the Twelfth International AAAI Conference on Web and Social Media* (ICWSM-18). AAAI Press.

Sanguinetti, M., Poletto, F., Bosco, C., Sanguinetti, V., Stranisci, M., & Russo, I. (2020). HaSpeeDe 2 @ EVALITA2020: Overview of the EVALITA 2020 hate speech detection task. *CEUR Workshop Proceedings*, 2765, 1–15.

Schmidt, A., & Wiegand, M. (2017). A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media* (pp. 1–10).

Selen, E. (2020). "The public immoralist": Discourses of queer subjectification in contemporary Turkey. *International Journal of Communication, 14,* 5518–5536.

Sharma, S., Agrawal, S., & Shrivastava, M. (2018). Degree-based classification of harmful speech using Twitter data. *arXiv:1806.04197.* https://arxiv.org/abs/1806.04197

Silva, L., Mondal, M., Correa, D., Benevenuto, F., & Weber, I. (2016). Analyzing the targets of hate in online social media. In *Proceedings of the Tenth International AAAI Conference on Web and Social Media (ICWSM 2016)* (pp. 687–690).

Simonsen, S. (2023). *Monitoring media pluralism in the digital era: Application of the Media Pluralism Monitor in the European Union, Albania, Montenegro, the Republic of North Macedonia, Serbia, and Turkey in the year 2022. Country report: Denmark.* Centre for Media Pluralism and Media Freedom.

Suryawanshi, S., Chakravarthi, B. R., Arcan, M., & Buitelaar, P. (2020). Multimodal meme dataset (MultiOFF) for identifying offensive content in image and text. *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying,* 32–41. https://www.aclweb.org/anthology/2020.trac-1.6/

Swamy, S. D., Jamatia, A., & Gambäck, B. (2020). Studying generalisability across abusive language detection datasets. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)* (pp. 940–950). Association for Computational Linguistics.

Ullmann, S., & Tomalin, M. (2020). Ethics and Information Technology, 22, 69–80. https://doi.org/10.1007/s10676-019-09516-z

Uludoğan, G., Dehghan, S., Arın, I., Erol, E., Yanıkoğlu, B., & Özgür, A. (2024, March). Overview of the hate speech detection in Turkish and Arabic tweets (HSD-2Lang) shared task at CASE 2024. In *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024)* (pp. 229–233). Springer. https://aclanthology.org/2024.case-1.32.pdf

UNHCR. (2024). *Refugees and asylum seekers in Turkey.*
https://www.unhcr.org/tr/en/refugees-and-asylum-seekers-in-turkey

United Nations. (2019). *United Nations strategy and plan of action on hate speech.* United Nations. https://www.un.org/en/genocideprevention/documents/advising-and-mobilizing/Action_plan_on_hate_speech_EN.pdf

Van Dijk, T. A. (2008). *Discourse and power*. Palgrave Macmillan.

Vidgen, B., & Derczynski, L. (2020). Directions in abusive language training data, a systematic review: Garbage in, garbage out. *Plos one,* 15(12), e0243300.

Warner, W., & Hirschberg, J. (2012). Detecting hate speech on the World Wide Web. In *Proceedings of the Second Workshop on Language in Social Media* (pp. 19–26).

Waseem, Z., & Hovy, D. (2016). Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop (NAACL SRW 2018),* 88–93.

Waseem, Z., Davidson, T., Warmsley, D., & Weber, I.. (2017). Understanding Abuse: A Typology of Abusive Language Detection Subtasks. In *Proceedings of the First Workshop on Abusive Language Online,* pages 78–84, Vancouver, BC, Canada. Association for Computational Linguistics.

Wiegand, M., Siegel, M., & Ruppenhofer, J. (2018). Overview of the germeval 2018 shared task on the identification of offensive language.

Wijaya, H. Y. (2022). Digital homophobia. Indonesia and the *Malay World,* 50(146), 52–72. https://doi.org/10.1080/13639811.2022.2010357

Williamson, M. (2023). A global analysis of transgender rights: Introducing the Trans Rights Indicator Project (TRIP). *Perspectives on Politics.* Advance online publication. https://doi.org/10.1017/S1537592723002827

Xue, L., Zhang, H., & Li, Y. (2016). The right to be forgotten: Users exploiting GDPR for hate speech removal. *International Journal of Information Management, 36*(3), 351–356.

Yuying, T., Nadine, N., Nadine, N., Mark, P., & Mark, P. (2013). Borders and margins: Giving voice to lesbian, gay, bisexual, and transgender immigrant experiences. *Journal of LGBT Issues in Counseling*. https://doi.org/10.1080/15538605.2013.785235

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019, June). SemEval-2019 Task 6: Identifying and categorizing offensive language in social media (OffensEval). *Proceedings of the 13th International Workshop on Semantic Evaluation,* 75–86. Association for Computational Linguistics.

Zampieri, M., Nakov, P., Rosenthal, S., Atanasova, P., Karadzhov, G., Mubarak, H., Derczynski, L., Pitenis, Z., & Çöltekin, Ç. (2020). SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). *arXiv preprint arXiv:2006.07235.*

Zhong, H., Li, H., Squicciarini, A., Rajtmajer, S., Griffin, C., Miller, D., & Caragea, C. (2015). Content-driven detection of cyberbullying on the Instagram social network. *International Joint Conference on Artificial Intelligence*, 3952–3958.

Zhuo, T. Y., Huang, Y., Chen, C., & Xing, Z. (2023). Red teaming ChatGPT via jailbreaking: Bias, robustness, reliability and toxicity. *arXiv*. https://doi.org/10.48550/arXiv.2301.12867

aS·u·lis

DISCOURSE
DIALOGUE
DEMOCRACY
LABORATORY

For your comments
and suggestions: