Hate speech operationalization: a preliminary examination of hate speech indicators and their structure

Jana Papcunová^{1,2} · Marcel Martončík³ · Denisa Fedáková¹ · Michal Kentoš¹ · Miroslava Bozogánová¹ · Ivan Srba⁴ · Robert Moro⁴ · Matúš Pikuliak⁴ · Marián Šimko⁴ · Matúš Adamkovič¹

Received: 25 February 2021 / Accepted: 21 September 2021
© The Author(s) 2021

Abstract

Hate speech should be tackled and prosecuted based on how it is operationalized. However, the existing theoretical definitions of hate speech are not sufficiently fleshed out or easily operable. To overcome this inadequacy, and with the help of interdisciplinary experts, we propose an empirical definition of hate speech by providing a list of 10 hate speech indicators and the rationale behind them (the indicators refer to specific, observable, and measurable characteristics that offer a practical definition of hate speech). A preliminary exploratory examination of the structure of hate speech, with the focus on comments related to migrants (one of the most reported grounds of hate speech), revealed that two indicators in particular, *denial of human rights* and *promoting violent behavior*, occupy a central role in the network of indicators. Furthermore, we discuss the practical implications of the proposed hate speech indicators—especially (semi-)automatic detection using the latest natural language processing (NLP) and machine learning (ML) methods. Having a set of quantifiable indicators could benefit researchers, human right activists, educators, analysts, and regulators by providing them with a pragmatic approach to hate speech assessment and detection.

Keywords Hate speech · Hate speech indicators · Hate speech structure · Migrants · Comments

Introduction

According to a Eurobarometer survey [1], 75% of people who follow or participate in online debates have witnessed or experienced abuse, threat, or hate speech. A more recent study by Eurostat [2] points to concerns about the online behaviors of children and young people and the harmful consequences—greater dependency, anxiety, and aggression. Both governmental and civil society organizations have put substantial effort into tackling hate speech (e.g., [3, 4, 20], etc.). Nonetheless, numerous scientific studies have indicated that the incidence of hate content is continually increasing [5–7].

✉ Jana Papcunová
jpapcunova@saske.sk

Marcel Martončík
martoncik@protonmail.ch

Denisa Fedáková
dfedak@saske.sk

Michal Kentoš
kentos@saske.sk

Miroslava Bozogánová
bozoganova@saske.sk

Ivan Srba
ivan.srba@kinit.sk

Robert Moro
robert.moro@kinit.sk

Matúš Pikuliak
matus.pikuliak@kinit.sk

Marián Šimko
marian.simko@kinit.sk

Matúš Adamkovič
matho.adamkovic@gmail.com

¹ Institute of Social Sciences, CSPS, Slovak Academy of Sciences, Košice, Slovakia

² Department of Psychology, Faculty of Arts, Pavol Jozef Šafárik University in Košice, Košice, Slovakia

³ Institute of Psychology, Faculty of Arts, University of Prešov, Prešov, Slovakia

⁴ Kempelen Institute of Intelligent Technologies, Bratislava, Slovakia

Successfully combating hate speech requires a proper multidisciplinary understanding of the phenomenon. Intuitively, people generally have some inner understanding of what hate speech is and when they cross the line. But the extent to which they assess the content and form can vary greatly [8]. Hate speech should be tackled and prosecuted, based on its *operationalization*. As stated by Brown [1, p. 45], “How can we ban it if nobody agrees about what it is?” A number of ad-hoc definitions have been introduced in recent years by legal entities, scholars, and social networks. Producing an accurate definition of online hate speech is, however, a challenging task given its multidimensional and complex nature. For example, legal definitions, which excessively concentrate on explicit hate, could make hate speech detection ineffective and targeted groups may feel that there is no one to protect them.⁹

Providing a *precise, universal, operable definition* of online hate speech that is generally accepted by multiple stakeholders—scholars and practitioners, who may come from different, unrelated fields—is even more challenging. It is nonetheless extremely important—such a definition would make it easier for stakeholders to share know-how, annotate data consistently, compare automatic detection methods, study, and compare the prevalence of hate speech across different platforms/countries, and would generally make communication easier, in the knowledge that others are likely to have some understanding of what we mean when we talk about online hate speech. Nevertheless, there is as yet no such theoretical definition [10, 11] and none is likely to appear and become widely accepted in the near future.

To address this problem and overcome the existing ambiguity, and lack of a unified and operational theoretical definition, we adopt a slightly different approach—we define hate speech empirically by providing a list of *hate speech indicators* (i.e., specific, observable, measurable characteristics that offer a practical means of defining the concept) and the rationale behind them. To the best of our knowledge, thus far such an approach has not been used to operationalize hate speech. Nevertheless, it has been used for other types of undesired user behavior and content (e.g., cyberbullying or fake news). In addition, some of the literature (e.g., [12, 13] already relies on some indicator precursors to overcome the ambiguity in definitions of hate speech and other relevant concepts.

The goal of this paper is therefore to operationalize hate speech using indicators and explore their use and the implications thereof, using a social science and computer science approach.

The main contributions of our approach are:

1. Instead of grounding hate speech in a psychological narrative, we take a pragmatic approach (inspired by diagnostic manuals like the Diagnostic and Statistical

Manual of Mental Disorders DSM-5 [14] or the International Classification of Diseases ICD-11 [15]) and (1) provide a list of hate speech indicators and the rationale behind them using a multidisciplinary approach, and (2) depict the structure of the indicators to reveal the core/peripheral indicators and the strength of the relationship between them.

2. Online hate speech is a research field that is open to diverse knowledge-based fields, and cross-disciplinary approaches are both key and often recommended [11, 16–18]. While social science researchers lack knowledge on automatic hate speech detection, most data analysts have no social science background. This paper takes into account the multiplicity of both computer science and social science approaches.
3. Although there are automatic methods of hate speech detection, using advanced NLP and ML methods, much could be done to aid social scientists interested in undertaking large-scale studies and in-depth research on hate speech. Having a set of quantifiable indicators could prove to be a beneficial pragmatic approach: (a) for measuring changes/assessments; (b) enabling international comparisons; (c) improving communication between scholars; (d) recognizing the extent of hate speech on the local/national level; (e) as a basis for legal proceedings; and from the computer science perspective, for (f) the annotation of extensive new hate speech datasets and for training novel explainable detection methods that provide explanations, or justifications, on what aspects of the text (e.g., a comment) lead to the decision of the automated method.

Background and related work

Definitions of hate speech

Definitions of hate speech can largely be found in four different contexts: (1) legal, (2) lexical, (3) scientific, and (4) practical; and these differ in scope and content. Below is a brief overview of each.

- (1) The purpose of legal definitions is straightforward: to identify messages that violate existing legal norms and require government regulation, i.e., messages that are publicly shared, incite, promote, or justify hatred, discrimination, or hostility toward a specific group and/or individual, based on certain attributes, such as race or ethnic origin, religion, disability, gender, age, sexual orientation/gender identity [4, 19–21]. There is no universally accepted legal definition of hate speech and different states, the primary duty bearers under international human rights, fall under different jurisdictions

- [22], Walker 1994 in Calvert [23]. What is considered “hateful” can be controversial and disputed. State actors, including Governments, legislatures, State authorities, and courts, therefore need a common understanding of the phenomenon before they can take relevant action to address it [24].
- (2) Five of the most well-known online dictionaries define hate speech as follows. Hate speech is: (1) “public speech that expresses hate or encourages violence towards a person or group based on something such as race, religion, sex, or sexual orientation” (The Cambridge Dictionary); (2) “speech expressing hatred of a particular group of people” (The Merriam Webster); (3) “speech that attacks a person or group on the basis of race, religion, gender, or sexual orientation” (Collins Dictionary); (4) “a statement expressing hatred for a particular group of people” (The MacMillan Dictionary); (5) “speech or writing that attacks or threatens a particular group of people, especially on the basis of race, religion or sexual orientation” (Oxford Learner’s Dictionary). All these definitions share the same characteristics. They define hate speech as a message (written or oral) that expresses hatred and encourages violence toward a specific group of people. These simple definitions concentrate on the meaning of the words that make up the collocation—hate and speech.
 - (3) Scientific definitions go much further, beyond the apparent meaning of the constituent words. The term hate speech is used in different disciplines such as economics, philosophy, sociology, psychology, or computer science, and, despite the lack of a definitional consensus, we can trace some of the common constituent characteristics. As Parekh [25] states, when defining hate speech, it is necessary to distinguish hate speech from related terms that do not fit the definition, such as expressing dislike, lack of respect, a demeaning view of others, disapproval, the use of abusive or insulting speech, and speech that does “not call for action” (p. 40). These subtle distinctions could be resolved by having a proper definition of hate speech.

According to Parekh [., pp. 40–41] hate speech “expresses, encourages, stirs up, or incites hatred against a group of individuals distinguished by a particular feature or set of features such as race, ethnicity, gender, religion, nationality, and sexual orientation. [...] and is often (but not necessary) expressed in offensive, angry, abusive, and insulting language”. Moreover Cohen-Almagor [25, 26] states that it is intended to dehumanize, harass, intimidate, debase, degrade, victimize, or incite brutality against the targeted groups. For an extensive philosophical analysis of the term hate speech, see Brown [9, 22].

Parekh [25] further recognizes three main aspects of the construct of hate speech: (1) it is directed at specific individuals or groups of individuals, (2) undesirable attributes are ascribed to the target group, thereby creating stigmatization; (3) it leads to discrimination because the ascription of undesirable attributes encourages the target to be seen as undesirable.

The third aspect of Parekh’s definition is highlighted by Gelber [27] as well—the inequality is a core element of hate speech. In the subordinate–superior relationship, hate speech puts the creator in a position of authority and the target in the position of the subordinate, which encourages the target to be seen as inferior and legitimizes discriminatory behavior [27].

According to Gelber [27], it is important also to note that the emotion of hatred does not have to be evoked for hate speech to cause harm. Hate is not and does not have to be a key element of the definition (see also [22], “myth of hate”). On the other hand, as Brown [9, 22] points out, if manifestations of hate speech did not include hate, then this concept would be grossly misleading, but it is not. In summary, a lot of speech evokes the strong emotion of hate, but not necessarily all of it.

Last but not least, it should be noted that hate speech need not relate only to a minority or disadvantaged social group as was previously thought by classic scholars such as Jacobs and Potter (1998) or Walker (1994) as cited in [28]. In past centuries hate speech was specifically directed at minorities (e.g., Afro-Americans, the Roma community).

In the computer science research literature (see our previous work [11] for an overview of the computer-science perspective on hate speech), the most common focus related to hate speech is its (semi-)automatic detection. With the help of natural language processing (NLP) and machine learning (ML), the models can classify a particular piece of content (usually textual) according to the binary presence of hate speech or a particular aspect thereof (e.g., whether it attacks an individual or a group), see the survey by Fortuna and Nunes [10] for more information. In computer science (and especially data science), there are two specific factors influencing the employed definitions of hate speech [11]: (1) adoption of how the data are already annotated (either by other researchers or by social media platforms, (2) the observable factors that can be identified from the content available in the dataset (i.e., without accessing more intangible clues such as the authors’ thinking, intent, or attitude that can be studied in other disciplines, e.g. psychology).

Like many other disciplines, computer science lacks a generally accepted definition of hate speech. In addition, it is not clear how hate speech relates to other similar/related concepts (e.g., toxic language) or superior concepts (e.g., abusive/offensive language) [11]. The boundaries between these concepts are often blurred and there are few works that

specifically tackle the distinctions. One example is Davidson et al. [29], who attempt to distinguish hate speech from other instances of offensive language in their work on the automated detection of hate speech.

- (4) Practical definitions can be found in the way online platforms (e.g., social media) and existing detection tools define hate speech. Facebook, Twitter, and YouTube are social media platforms that mediate online communication and have developed their own definitions of hate speech to ensure users stick to the rules and as part of their internal regulatory policies (e.g., terms of service or community standards). They have signed a Code of Conduct on the regulation of illegal hate speech with the European Commission [20]. Facebook's Community Standards [30] define hate speech as "a direct attack against people on the basis of what they call protected characteristics: race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity and serious disease".

The practical tools for identifying and moderating hate speech, such as profanity filters, content moderation filters, or human driven approaches, have been widely studied [5, 10, 11, 31, 32]. In the context of social media platforms, the findings on utilization of practical tools are somewhat puzzling as they either show that social platforms perform too much or too little content moderation and lack transparent decision-making processes [33].

Capturing the complexity of hate speech in a unified, precise, and operable definition (either theoretical or more empirically oriented) could potentially lead to a better understanding of hate speech and help reduce its occurrence online. Fortuna and Nunes [10] reviewed all four types of (legal, lexical, scientific, and practical) hate speech definitions. Following a content analysis, they proposed their own theoretical definition (probably one of the most fleshed out ones) [10, p. 5]:

Hate speech is language that attacks or diminishes, that incites violence or hate against groups, based on specific characteristics such as physical appearance, religion, descent, national or ethnic origin, sexual orientation, gender identity or other, and it can occur with different linguistic styles, even in subtle forms or when humour is used.

There are other useful and inspiring attempts at a universal definition reached through a process of characterizing the toxicity of comments [34, 35], or utilizing context information [36, 37].

Nevertheless, when it comes to their practical application (e.g., during data annotation for machine learning), these theoretical definitions are still too vague and difficult to capture.

In addition, many local factors (e.g., cultural, platform, task at hand) make it difficult or even impossible to produce a universal theoretical definition. Indeed, we argue that the potential for finding one seems unlikely at the moment.

Consequently, our aim in this paper is to define hate speech empirically by identifying specific, measurable, observable hate speech indicators.

Indicators of hate speech

An indicator is "a sign that shows you what something is like or how a situation is changing" (Oxford Learner's Dictionary). Previous studies have looked at indicators of cyberbullying [38], operational indicators of fake news [39], and behavioral indicators related to sexting [40]. Indicators are also frequently used in psychopathology as a means of conceptualizing and diagnosing mental health problems. We propose that operationalizing hate speech in the form of indicators could fulfil the same purpose—provide objective, measurable cues to help "diagnose" hate speech and find effective interventions.

Some existing research papers on hate speech deal with the content characteristics that describe hate speech and can be considered precursors to hate speech indicators. However, these are not primarily aimed at identifying hate speech indicators using an adequate methodology and so do not assess them satisfactory.

First, Fortuna and Nunes [10] identified four dimensions (that can be considered analogous to indicators) that allowed them to compare definitions of hate speech: (1) hate speech has specific targets, (2) hate speech is to incite violence or hate; (3) hate speech is to attack or diminish; and (4) hate speech can be expressed through (and hidden in) humor.

Secondly, some computer science studies have gone further than producing a binary classification and have distinguished different aspects of abusive/offensive language (as a superior concept to hate speech). Waseem et al. [12] proposed a typology of abusive language based on defining two dimensions: (1) directed versus generalized, and (2) explicit versus implicit abusive language. Zampieri et al. [13] similarly annotated and automatically classified offensive language according to directness (targeted insult, untargeted) and target identification (individual, group, other). By emphasizing these critical aspects of offensive/abusive language, these studies reduce the ambiguity in definitions of hate speech and related concepts. Ousidhoum et al. [41] focused specifically on hate speech and manually annotated a multilingual dataset of Tweets in three languages, distinguishing five aspects: directness, hostility type (e.g., abusive, hateful, offensive), target attribute (i.e., what is being targeted, e.g., origin, gender, religion), target group (i.e., who is being targeted, e.g., individual, women), and annotator sentiment (i.e., the emo-

tion felt by the annotator after reading the Tweet, e.g., disgust, fear, anger).

Other things resembling indicators can be identified in data annotation methodologies as well. The manual annotation of hate speech is a subjective process and annotators need extensive cultural and societal knowledge [42]. Human annotators must therefore be provided with detailed instructions on how to recognize and label hate speech consistently. For this purpose, Waseem and Hovy [43] proposed a list of 11 criteria to guide annotators, examples items are: hate speech (1) uses sexist or racial slurs, (2) criticizes a minority and uses straw man arguments; and (3) defends xenophobia or sexism.

Finally, in NLP and ML models, training of detection models commonly involves a feature engineering step. Features utilized to train models (especially those with a strong prediction capability) can be considered a kind of indicator as well. Fortuna and Nunes [10] analyzed a number of detection methods and identified two categories of features: general features commonly used in text classification approaches and specific hate speech detection features. The latter category, which is more closely related to the concept of indicators, was further divided into several subcategories of features that are intrinsically related to the characteristics of hate speech: (1) othering language, (2) perpetrator characteristics, (3) objectivity-subjectivity of the language, (4) declarations of superiority of the ingroup, (5) a focus on particular stereotypes, and (6) intersectionism of oppression.

To conclude, despite the increasing understanding in the social sciences of the factors that may underlie online hateful behavior [44, 45], we found no studies attempting to systematically develop a set of indicators covering the main aspects of hate speech. To fill this gap and conduct a systematic study of hate speech indicators, we aim to provide an empirical definition of hate speech, while taking into account both social science and computer science approaches (with the help of existing research and interdisciplinary experts).

Choosing the hate speech indicators

The primary goal of our research was to propose an empirical definition of hate speech. As stated by Chaffee [46] “an empirical definition of a concept enables us to conclude whether an event we encounter is an instance of that concept or not.” This social science insight has extremely important relevance for computer science and supervised machine learning—in practice, it could improve the accuracy and effectiveness of hate speech detection and removal.

The concept of hate speech is a social construct. It represents agreement between people about what speech is hateful and what is not. Therefore, we started our search for an empirical definition by holding focus group discus-

sions with two groups: group 1 comprising 12 participants selected and assembled by the authors of this manuscript; group 2 consists of a team of researchers—psychologists with different levels of seniority—primarily the authors of this manuscript. The participants (group 1), who varied in gender, age and educational background, discussed and commented on the concept of hate speech, based on their personal experiences. Research ethics were concerned with respecting the research participants throughout the study. The focus groups of researchers took the form of a structured informal discussion described below. Eight focus group sessions were conducted over two months, April–May 2020, with the team of researchers (group 2) in one-hour regular discussion sessions. All eight focus groups were conducted utilizing a dual moderator model [47, 48] until the mutual agreement on the selected indicators was reached. First the discussion focused on hate speech indicators in general, but as the sessions continued the focus shifted more towards the hate speech indicators relating to migrants. The focus on migrants was selected on the basis of current reports by: (1) the European Commission [49] stating that xenophobia, anti-migrant hatred (15%), and anti-gypsyism (9.9%) are among the most reported grounds for hate speech; (2) UNHCR [50] about powerful narratives denigrating refugees and turning them into objects of fear; (3) the ECRI report on Slovakia [51] pointing to escalation in hate speech against Jews, Muslims, migrants, Roma, and black people.

The most significant outcome of the focus group discussions and subsequent non-systematic review of the hate speech literature was the creation of a list of 14 hate speech indicators relating to migration and migrants (see Fig. 1). The indicators cover the various features of hate speech in a form that is easily accessible to policy makers, educators, researchers, and human rights activists. Crucially, the occurrence of each indicator is easily quantifiable.

The list of 14 indicators (including general information about the purpose of the study and a description of each indicator including a representative example) was then sent to experts who were asked to evaluate the content validity of the indicators and for any additional comments. The experts were all based in Slovakia and specialized in asylum and migration law (Human Rights League), sociology (Centre for the Research of Ethnicity and Culture), leading media (*SME*¹), a non-profit organization (DigiQ²), IT, social work and human rights protection, and psychology. Representatives of minorities in Slovakia (Islamic Foundation Slovakia)

¹ *Sme*, in full *Denník SME (We Daily Newspaper)*, is one of the most widely read broadsheets in Slovakia. The website, sme.sk, is one of the most visited internet portals in Slovakia.

² DigiQ is an NGO that promotes digital citizenship, cybersecurity education, and operates a hate speech phone line. It is also a trusted flagger for Facebook, YouTube, Instagram, TikTok, and Snapchat.

Fig. 1 Initial 14 hate speech indicators identified via the literature review and focus group discussions

were also sent a copy. An online consultation was held with one of the experts in relation to her comments and recommendations. Altogether, 18 experts responded and evaluated the validity of the indicators. Their task was to indicate (on a 5-point scale where 0 = does not indicate hate speech at all; 4 = fully indicates hate speech) the intensity of each indicator. The intraclass correlation coefficient (ICC; indicates the level of agreement between the raters) shows very good agreement between the experts' evaluations, $ICC = 0.80$, $p < 0.001$. The experts concluded that denial of other rights is the most intensive hate speech indicator whereas the weakest hate speech indicators were manipulation, slant, and use of tendentious analogies (for more details see <https://osf.io/mctkv/>). After checking the descriptives of the expert ratings, we further discussed and re-evaluated the content of the indicators and merged those we deemed very similar, reducing the list to the final 10 indicators. For an overview of the process see Fig. 2.

List of hate speech indicators and the rationale

The list of 10 indicators and the rationale behind them was arrived at through a process of explication, identifying the theoretical and empirical definitions of hate speech and then evaluating and modifying them.

Below we provide a description of each indicator. The purpose of the indicators is to provide a simple but comprehensive picture of the hate speech phenomenon, which can then be used to help to identify, evaluate, or summarize the results achieved along the way in tackling hate speech, and may help in monitoring hate speech trends in different regions and cultures. The indicators can also be used for setting goals and benchmark criteria. The indicators are relevant, unambiguous, comprehensible, interpretable, and comparable over time.

1. *Sexist language*: sexist hate speech relates to utterances which spread, incite, promote, or justify hatred based on sex, gender, or sexual preference, with the aim of humiliating and objectifying the target, destroying their reputation, and making them vulnerable and fearful. It also includes second sexism—discrimination against men and boys. The feminist aspect is even more prominent in relation to migration. Almost half of migrants are women and girls (Gender Statistics [52]). Negative depictions often appear in the media, where female migrants often face double discrimination—as women and as migrants.
2. *Attacking minorities as a “traditionally disadvantaged group”*: this indicator includes any sort of racial offense, xenophobia, antisemitism, discrimination on the grounds of the person's age, disability, ethnicity, national or reli-

Fig. 2 Flowchart of the process of selecting the indicators

gious group aimed at inciting intolerance or racial and ethnic hatred.

5. *Denial of fundamental human rights*: this includes any calls for the exclusion or segregation of a person or group of people, being denied the right to promote one's language and traditions, peaceful assembly etc.
6. *Promoting violent behavior*: comments aimed at inciting others to commit acts of violence or that can reasonably be expected to have that effect. This also includes excusing violence, promoting negationism, and condoning terrorism.
7. *Problematic hashtags, nicknames, symbols*: this can also mean referring to an organization that has committed hate crimes.
8. *Ad hominem attacks*: these are personal attacks on a person's character or motive, based on feelings of prejudice, rather than facts, logic. They include

accusations of lying, ignorance, stupidity, and condescension.

9. *Negative stereotypes of a minority*: these are negative expectations about the out-group, and appear alongside negative emotions (e.g., fear, anger) towards the out-group. They promote the view that the targeted group is inferior or subordinate and are closely linked and related to Indicator no. 2.
10. *Texts containing ambiguous statements, irony, sarcasm*: making generalized negative statements about minority groups, ridicule, intimidation, hostility. Mocking the concept, events or victims of hate crimes, even if no real person is discussed.
11. *Manipulative texts/misinterpretation of the truth*: any misinterpretations made purposefully, with the intention of fooling or harming someone. For instance, denial of historical events.
12. *Slurs and vulgarisms*: direct, explicit verbal or non-verbal attacks, using insulting labels to refer to an

individual or group based on their race, ethnicity, national origin, religious affiliation, or other attribute.

Our proposed list of indicators is similar to work by Waseem and Hovy [43] based on critical race theory. They provided a list of criteria for human annotators identifying and annotating hate speech based on Twitter feeds. It differs from our indicators in that we relied on the views of the multi-disciplinary experts involved (i.e., we did not base our list on theory alone) and, while some of the indicators are the same (e.g., hate speech is the use of a sexist or racial slur, attacks a minority, uses problematic hashtags, negatively stereotypes a minority), others are slightly different. In addition, following the preliminary exploratory examination of the structure of hate speech (cf. “Using indicators to examine the structure of hate speech”), it seems that the denial of human rights and promotion of violent behavior (not present in the work of Waseem and Hovy [43]) plays a central role in the network of indicators.

Using indicators to examine the structure of hate speech

To obtain better insights into hate speech, we examined the internal structure and dynamics of hate speech indicators via a network analysis. In recent years, the network approach to the conceptualization of social and behavioral phenomena has appeared in the behavioral sciences and has become increasingly popular, especially in psychopathology (e.g., [53]). The network approach shifts the focus away from latent constructs (i.e., unobservable entities that cause a set of observable behaviors) and toward complex systems for grouping variables. In this perspective, a phenomenon occurs owing to a causal system of mutually interacting variables. The four network theory principles for conceptualizing mental disorders have been summarized by Borsboom [54] and can be applied to hate speech contexts and the indicators we propose. That is, hate speech can be regarded as a specific combination of indicators of varying intensities that mutually interact and reinforce/weaken each other (the indicators theoretically correspond to the nature of the construct, i.e., the indicators proposed above).

The full procedure of hate speech structure examination was as follows:

1. A sample of around 240 comments about migration/migrants was obtained via a qualitative analysis of public social media and local news (Facebook, *SME*,

HNonline,³ *Magazin*⁴) or created by some of the members of the research team. The comments differed in the level of hate expressed: some slightly misinterpreted historical facts, whereas others contained vulgarisms or explicitly called for violent acts. To prevent the range restriction phenomenon, neutral comments were also selected to capture the whole spectrum of speech (ranging from neutral comments to very hateful comments).

2. Members of the research team were then asked to rate each comment in terms of the indicators described above and to assess the hate speech in the comment overall. Each comment was independently rated by four members of the research team, consisting primarily of the authors of this manuscript.
3. The process of selecting comments was stopped once to check how well the comments represented the whole spectrum of hate speech severity (based on the ratings). Sexism, irony, and manipulative texts from the spectrum of positive, neutral, and very hateful comments seemed to be underrepresented, so we added another 60 comments.
4. Before proceeding to the analysis, we screened the comments one last time and excluded those ($N = 13$) not explicitly related to migration. The resulting dataset consisted of 283 representative comments. The final set of comments (in Slovak) is available at <https://osf.io/452pb/>.
5. Once the selection/extraction of the comments had been finalized, the level of agreement between the raters for each indicator was calculated. Despite all the ICCs being significant, the average level of agreement was $ICC = 0.64$, indicating moderate reliability, according to the rule of thumb. However, given that some indicators, like ad hominem attacks, manipulative statements, or ambiguous or ironic statements, are naturally dependent on the person interpreting them, rater agreement can in fact be better than appears at first glance.
6. The rating for each indicator was then averaged (individual ratings can be found in the dataset at <https://osf.io/g3b87/>).
7. Furthermore, we checked how well the average perception of hate speech correlated with the general impression of hate speech, the average severity of the indicators and their product (the sum score of the indicators), as well as how strongly the indicators were correlated. Excluding sexism, there was a moderate to strong relationship between all the indicators and the general impression of hate speech. All the indicators correlated positively, with the magnitude of the correlations ranging from 0.02 to

³ *Hospodárske noviny* (abbreviated as HN; Economic Newspaper) is a daily economic newspaper published in Slovakia.

⁴ Political and social magazine run by Slovak politician and leader of a far-right neo-Nazi political party called the Kotlebaists.

0.72; the average correlation between the indicators was $r = 0.32$.

8. A network analysis approach was used to examine the structure of the hate speech indicators. It tells us how the indicators (nodes) of a phenomenon interconnect (edges), which play a more central role (strength and expected influence: how strongly an indicator is related to other indicators; closeness and betweenness: how well an indicator connects with other indicators), and which do not. For the purposes of this demonstration, a network consisting of the average evaluation of each of the hate speech indicators was computed. Here we depict the structure of hate speech (only for the 283 comments which featured at least one hate speech indicator; the network of all the comments is available in the supplementary materials), show the estimates of the centrality measures, and very briefly discuss the stability of the network. For more details on the technical side, please see the supplementary materials at <https://osf.io/g3b87/>.

For descriptive purposes, the correlation matrix and the corresponding network of binary relationships are presented in Table 1 and Fig. 3a respectively. A regularized network of the hate speech indicators is depicted in Fig. 3b. Please note that Fig. 3a, b are visualizations of the networks—although these are intuitive to comprehend, visualizations can be misleading and so interpretations should be based on the statistical estimates. Regarding the estimated centrality measures (Fig. 4), the core indicator of hate speech appears to be: (a) *denial of fundamental human rights* accompanied by (b) *promoting violent and aggressive behavior*, and (c) *the use of vulgarisms*. In contrast, a rather peripheral role in the network is played by the use of: (a) *ambiguous statements, irony, or sarcasm* and (b) *the use of problematic hashtags or nicknames*. Paradoxically, after controlling for the other variables in the network, some of the correlations between denial of human rights and the other indicators ended up negative (that is why the expected influence drops so substantially compared to the strength estimate).

The results are, to our knowledge, the first attempt to represent the structure of hate speech and as such, should be interpreted with caution. The attribute of promoting violent and aggressive behavior features in both legal and scientific definitions. This indicator seems to distinguish hate speech from distinct but related behaviors, such as dislike, lack of respect, or disapproval since, as Parekh [25] states, it “calls for action”. Our analysis shows that insulting and abusive language, previously considered an important but unnecessary component by scholars, seems to be another core indicator of hate speech. Since both indicators are easy to operationalize, they provide clearer guidance for the effective identification of hate speech. Some scholars [25, 27] consider inequality or the subordinate-superior relationship to be a defined com-

ponent of hate speech. We suggest that this vaguely defined component manifests in specific types of speech as the third core indicator, denial of fundamental human rights. Denial of fundamental human rights places the person using hate speech in the position of a superior who determines who should and should not be considered a full-fledged human being. Again, focusing on the list of fundamental human rights and manifestations of their denial could facilitate the identification (in a public context) and measurement (in an academic context) of hate.

Practical implications

This paper presents a new approach to understanding hate speech online. Although there are automatic ways of detecting hate speech using advanced ML and NLP methods, still much could be done to aid social scientists interested in under-taking large-scale studies and in-depth research on hate speech. Our research into the investigation of hate speech indicators and their use in combating online hate speech incorporates a social science and computer science perspective, while offering some practical implications.

Having a set of quantifiable indicators could benefit researchers, human right activists, educators, analysts, and regulators by providing them with a pragmatic approach to hate speech assessment and detection. This pragmatic approach could benefit:

1. Scholars and researchers seeking to improve the way they communicate about the phenomenon of online hate speech. Without a clear concept and understanding of the concept, how can scholars and researchers know what speech should be targeted?
2. The non-profit sector, when conducting comparative international studies, measuring changes/assessments in the local community, or fostering debates, discussions, or the exchanging of ideas on how to protect disadvantaged client groups.
3. Regulators who need to identify the extent of hate speech on the local/national level, provide recommendations and mechanisms for fighting against and preventing hate speech, and for facilitating further efforts and initiatives.
4. Social media moderators and online portals who could use it as a basis for hate speech detection and moderation and cultivating discussion.
5. Legal proceedings. As Pejchal [55] stresses in a study analyzing legal instruments in post-communist Slovakia and the Czech Republic, there is a need to support and create a favorable environment for discussing democracy and the values it stands for—free speech versus hate speech.

Table 1 Correlation matrix of the hate speech indicators, general hate speech impression, and a sum score of the indicators

	1	2	3	4	5	6	7	8	9	10	11	12
1	–											
2	0.25	–										
3	0.13	0.45	–									
4	0.17	0.66	0.72	–								
5	0.14	0.37	0.32	0.55	–							
6	0.14	0.43	0.48	0.55	0.48	–						
7	0.20	0.73	0.31	0.54	0.28	0.32	–					
8	0.34	0.25	0.16	0.27	0.19	0.17	0.18	–				
9	0.04	0.35	0.13	0.26	0.09	0.18	0.46	0.13	–			
10	0.14	0.47	0.40	0.61	0.51	0.55	0.36	0.13	0.07	–		
11	0.29	0.83	0.60	0.84	0.52	0.54	0.72	0.38	0.38	0.60	–	
12	0.38	0.81	0.65	0.87	0.62	0.65	0.73	0.42	0.42	0.70	0.92	–

Note: 1: sexism, 2: attacks on minorities, 3: denial of human rights, 4: promoting violent behavior, 5: problematic hashtags and nicknames, 6: ad hominem attacks, 7: negative stereotypes, 8: ambiguous statements and irony, 9: manipulative text, 10: slurs and vulgarisms, 11: general hate speech impression, 12: hate speech score (sum of numbers 1–10)

Figure 3a. A network of binary correlations between the hate speech indicators

Figure 3b. A regularized network of partialized correlations between the hate speech indicators

1: Sexism, 2: Attacks on minorities, 3: Denial of human rights, 4: Promoting violent behavior, 5: Problematic hashtags and nicks, 6: Ad hominem attacks, 7: Negative stereotypes, 8: Ambiguous statements and irony, 9: Manipulative text, 10: Slurs and vulgarisms

Fig. 3 Visualization of hate speech structure

While the proposed indicators may prove beneficial in the manual detection of hate speech in the above-mentioned scenarios, their greatest potential lies in automatic hate speech detection. First, many existing datasets (e.g., [13, 29, 56]) contain a label that a post is an instance of hate speech or offensive language, but often no finer-grained description or explanation is given. Although the annotators had definitions or guidelines to follow (see, e.g., [43]), these are not captured in the data. Thus, the same label in two different

datasets may have a different meaning, which can hamper the generalizability of the trained models and prevent the use of datasets for transfer learning or even a simple comparison of the different models' performances. Having a dataset annotated with the proposed indicators would alleviate these issues.

Moreover, the indicators can be used as input features or weak labels to train hate speech detection models. Such models would have the added benefit of explainability (e.g., the

Fig. 4 Centrality measures of a regularized network of the hate speech indicators (ordered by strength). Note: The x-axis represents the standardized score for each variable (y-axis) in the index. A basic (simplified) description of the indices is presented in the body of the paper

above (The structure of hate speech, point no. 8). For more detailed information on the strength, closeness, and betweenness please see the tutorial paper by Constantini et al. [77], for more information on expected influence please see [78]

comment is labeled as hate speech because it promotes violent behavior and negatively stereotypes a minority), which is important for moderating social media content.

However, for the indicators to be used as weak labels or as explanations, they need to be automatically quantifiable from an input text (e.g., a social media post). In Table 2 we show how the proposed indicators occur in existing datasets or lexicons, which could be used to train models to automatically detect or quantify them. Given the available datasets and the current level of maturity of existing NLP approaches, the proposed indicators vary in readiness for automatic detection. *Problematic hashtags, nicknames and symbols* as well as *slurs and vulgarisms* can be detected with a high level of precision using simple word lexicons. However, the cultural sensitivity of certain word usages (words that are problematic in some cultures, but not in others) remains an open problem. Curated lexicons of problematic hashtags are still missing, although existing lists of hate speech words⁵ could be used for this purpose. Language aimed at the *denial of fundamental human rights* or *promoting violent behavior* can be detected with machine learning methods based on sufficiently large datasets, since it is often characterized by a relatively homogenous set of features. While existing datasets often contain a label pertaining to violence and aggressiveness [57–60], the denial of fundamental human rights is rarely explicitly labeled, despite our analysis showing it is a strong indicator of hate speech.

Sexist language or *language attacking minorities as a “traditionally disadvantaged group”* can also be detected with machine learning methods and there are datasets available containing these labels (e.g., [41, 43, 57, 61]). However, certain types of these are still hard to detect, e.g. latent sexist language, ambiguous statements or language drawing on a broader context and not just individual phrases or sentences. *Stereotypical language* can be detected with machine learning methods as well, but it is explicitly labeled in only a few datasets [59, 60, 62]. Another issue is the sheer amount of different stereotypes in different cultures.

Certain types of *ad hominem* attacks are relatively easy to detect, especially attacks in which other indicators feature, such as slurs and stereotypical language. The existing datasets that contain name calling or smears as a type of *ad hominem* reflect this [63, 64]. But there are context-sensitive *ad hominem* attacks whose detection requires a natural language understanding level not yet achieved by current systems. For instance, “He spends all his time in a library” might be an *ad hominem* attack aiming at a person’s practical skills, but in other contexts it could in fact be praise.

Despite the existing research, automatic *irony* or *sarcasm* detection or automatic detection of a *manipulative style* or *misinterpretations* is still far from a solved problem. Although there are labeled datasets on irony [65], humor [66], and manipulative (propaganda) techniques [63, 64], the current systems do not possess a level of natural language understanding that would enable them to reliably detect such phenomena. Yet our analysis shows that the use of ambigu-

⁵ For example: <https://hatebase.org/>.

Table 2 Map of proposed indicators to existing NLP datasets on hate speech and related phenomena, such as offensive and abusive language and the assessment of the current level of readiness for the automatic

detection of each indicator given the availability of data and the capability of existing approaches

Indicator	Occurs in existing datasets	Level of readiness for automatic detection
Sexist language	[41–43, 57, 61, 73, 74] (Italian, Spanish)	Sexist language can to some extent be detected using machine learning methods and large enough datasets. However, certain types of sexist language are still hard to detect, e.g. latent sexist language, ambiguous statements or sexist language that draws on a broader context and not individual phrases or sentences
Attacking minorities as a “traditionally disadvantaged group”	[42] (racism), [43] (racism), [41, 57] (labeled as origin)	Language attacking minorities can be detected using machine learning methods and large enough datasets, since it is often characterized by a relatively homogenous set of features. However, certain types of attack are still hard to detect, e.g. ambiguous statements or attacks that draw on a broader context and not individual phrases or sentences
Denial of fundamental human rights	Partially in [60] (Italian language dataset) as intensity label	Language containing the denial of fundamental human rights can be detected using machine learning methods and large enough datasets, since it is often characterized by a relatively homogenous set of features
Promoting violent behavior	Coded as aggressiveness in [57–60] (last two in Italian) Coded also as intensity in [60] (in Italian)	Language promoting violent behavior can be detected using machine learning methods and large enough datasets, since it is often characterized by a relatively homogenous set of features
Problematic hashtags, nicknames, symbols	As far as we know, there is no curated list of problematic hashtags However, there is a list of hate speech vocabulary, which could be used for this purpose: https://hatebase.org/	Problematic hashtags, nicknames and symbols can be detected with a high level of precision using simple word lexicons. Issues with the culturally sensitive usage of certain words (words that are problematic in some cultures, but not in others) remain an open problem
Ad hominem attacks	[75] Name calling label in [63] Name calling and Smears labels in [64]	Certain types of <i>ad hominem</i> attacks can be easy to detect, especially attacks where other indicators are present, such as slurs, stereotypical language. But there are context-sensitive <i>ad hominem</i> attacks whose detection requires a natural language understanding level not yet achieved by the current systems. E.g. “He spends all his time in a library” might be an <i>ad hominem</i> attack aiming at a person’s lack of practical skills, but in other contexts it could in fact be praise
Negative stereotypes of a minority	Explicitly present only in Italian language datasets: [59, 60, 62]	Stereotypical language can be detected using machine learning methods and large enough datasets, since it is often characterized by a relatively homogenous set of features (e.g. specific words and phrases). Issues remain with the sheer amount of different stereotypes that exist in different cultures
Texts with ambiguous statements, irony, sarcasm	Irony: [65] Humor and offense: [66]	Despite the existing research, automatic irony or sarcasm detection is still a challenge. Current systems do not possess the level of natural language understanding to reliably detect such phenomena
Manipulative texts/misinterpretation of the truth	[63, 64]	Despite the existing research, automatically detecting a manipulative style or misinterpretations is still a challenge. Current systems do not possess the level of natural language understanding to reliably detect such phenomena
Slurs and vulgarisms	Profanity filter, e.g. https://github.com/rominf/profanity-filter/blob/master/profanity_filter/data/en_profane_words.txt List of ethnic slurs: https://en.wikipedia.org/wiki/List_of_ethnic_slurs There is also a list of hate speech vocabulary, which might be used for this purpose: https://hatebase.org/ Offensive language datasets, e.g., [13, 29, 76]	Slurs and vulgarisms can be detected with a high level of precision using simple word lexicons. Issues with the culturally sensitive usage of certain words (words that are slurs in some cultures, but not in others) remain an open problem

ous statements, irony or sarcasm is a peripheral indicator of hate speech.

Limitations of the present study and suggestions for future research

The present study has several limitations that can be divided into two main areas—robustness of procedure and technicalities related to the analysis. The list of indicators is probably not complete, and one can expect it to change as more ideas or empirical evidence come to light. This is not surprising given that even relatively well-established concepts in diagnostic manuals (e.g., PTSD, compare DSM-IV and the DSM-5) are continually evolving. It could be argued that the procedure employed here relied heavily on experts' judgement (both in the selection of the indicators as well as their evaluation). Although expert judgement is essential at all stages of the scientific process [67], there is some evidence (e.g., [68]) that it may be less accurate than intuitively expected. To prevent systematic mistakes in the conceptualization of a phenomenon, it would perhaps be better to use a pragmatic data-driven approach (e.g., [69]) combined with an expert consensus. This could be a future area of research. Furthermore, the validity and hence generalizability of the set of indicators is likely to be dependent on the way the topic (e.g., LGBTI, gender, political issues) interacts with the cultural setting. To tackle this issue, replication studies on different topics in different cultures would be a very welcome addition. Applied to the present research, the results are probably limited by the specifics of Slovak culture/language and the polarizing topic of migration which may have distorted the ratings.

From the technical perspective, there are two major limitations to this study. First, the input number for calculating the network is taken from the averaged ratings of the hate speech indicators obtained from a fixed set of independent raters. Although this is a legitimate practice [70], a larger pool of expert raters would make the results more robust. Second, given the small sample (i.e., the number of comments), the stability of the observed estimates and the robustness of the whole network is moderate at best (for the results of simulations see <https://osf.io/g3b87/>). As noted above, the present network primarily serves as a probe into the topic and much more extensive research and replication is needed in order to obtain credible information about the structure of hate speech.

From the computer science perspective, the labeled dataset of hate speech comments is too small to be used in training automated detection models. However, this was beyond the scope of this work and a collection of a larger dataset is planned in the future. One limitation for future research and, more importantly, for practical deployment is that, as

discussed above, not all of the proposed indicators are easily quantifiable (detected) by automatic means using ML and NLP methods (given the current state-of-the-art and available resources). However, the indicators, identified in our analysis as most significant for the operationalization of hate speech, can be detected with machine learning methods given large enough datasets, since they are characterized using a relatively homogeneous set of features. Lastly, one could object that our proposed indicators replace a simple binary classification of whether a comment is or is not hate speech with a set of classification/regression problems. But as argued above, the main advantage is that using indicators leads to unambiguous, interpretable, and comparable labels, which can then be used to detect and explain hate speech. Also, as shown above, most indicators can already be found in some form or other in the existing datasets, which points to their relevance and feasibility.

Conclusion

Hate speech is prevalent in all societies, across various social media platforms, and continues to present a challenge to practitioners and regulators. The latest report from the Organization for Security and Co-operation in Europe OSCE [71] includes hate crimes and bias-motivated crimes but fails to provide information about discrimination or hate speech owing to the lack of consensus on whether these acts should be criminalized. The more vague a theory is, the less it helps us describe, explain, predict, and control a phenomenon [72]. This applies to weak, narrative theories, which can be contrasted with strong theories that enable phenomena to be precisely formalized [72]. A good hate speech theory would be useful in providing us with a recipe for accurately identifying forms of hate speech. Our goal was to operationalize hate speech by proposing indicators and examining their structure. We defined hate speech in the context of migrants as any text that promotes *violent behavior, denies human rights, contains slurs, vulgarisms, or ad hominem attacks, uses negative stereotypes, or purposefully manipulates the truth or historical facts*. The indicators are interchangeable, and may not all be present, but we consider the promotion of violent behavior, *use of vulgarisms, and denial of human rights* to be the core aspects of hate speech. The proposed indicators can be directly used in practice (i.e., a more precise assessment of hate speech severity by rating the various indicators) and further developed by other researchers. The grey zone deserves special attention and, from the pragmatic perspective, would benefit from a more nuanced conceptualization of the hate speech phenomenon.

For future research, if we are to fully grasp the phenomenon, resources must be shared between researchers, and studies should be conducted/replicated across different

cultures and topics, as there is a pressing need for more evidence-based information. In this regard, we plan to use the proposed methodology and indicators to collect a larger sample of labeled comments, of such a volume that it can also be used to train automatic hate speech detection models. To the best of our knowledge, our work provides the first (albeit small) public dataset of labeled hate speech comments in the Slovak language, and the extended dataset will be a significant contribution to the limited NLP resources available in Slovak (and other Slavic languages for that matter). Lastly, we plan to examine the methods for the automatic detection (or quantification) of the proposed indicators using the available resources and state-of-the-art NLP approaches as discussed in this paper.

Acknowledgements We would like to thank Maria Bielikova for the contribution to the paper and all additional experts and colleagues who provided feedback and/or helped with the preparation of the materials. Thank you for your valuable input. We also would like to express our sincere gratitude to anonymous reviewers for their valuable feedback on the manuscript.

Author contributions (CRediT taxonomy): JP: conceptualization, writing—original draft, investigation, writing—review and editing; MM: writing—original draft, writing—review and editing; DF: conceptualization, supervision, writing—original draft. MK: conceptualization, investigation, data curation; MB: conceptualization, writing—review and editing; MA: conceptualization, formal analysis, methodology, writing—original draft, writing—review and editing; IS: conceptualization, writing—original draft, writing—review and editing; RM: conceptualization, writing—original draft, writing—review and editing; MP: writing—original draft, writing—review and editing; MS: conceptualization, writing—review and editing. All the authors have revised the manuscript, made a substantial contribution to the work, and approved it for publication.

Funding This study was supported by the SAS PhD Student Program Grant under Contract no. APP0175, the Slovak Research and Development Agency, under Contract nos. APVV-18-0140 and APVV-20-0319, and VEGA 2/0068/19 and by TAILOR, a project funded by EU Horizon 2020 research and innovation programme under GA No. 952215.

Availability of data and materials The materials, data, and analytic code are freely available at <https://osf.io/9fykd/>.

Declarations

Conflict of interest None declared.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the

permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. EU Commission (2016) Eurobarometer, S. 452, media pluralism and democracy. Retrieved from <https://ec.europa.eu>
2. Eurostat (2020) Being young in Europe today—digital world. Retrieved from <https://ec.europa.eu/eurostat>
3. Council of Europe (2013) No hate speech movement. Retrieved from <https://www.coe.int/en/web/no-hate-campaign>
4. ECRI (2016) General Policy Recommendation No. 15 on combating hate speech. Retrieved from <https://coe.int>
5. Mathew B, Saha P, Tharad H, Rajgaria S, Singhanian P, Maity SK, Mukherjee A et al (2019) Thou shalt not hate: countering online hate speech. In: Proceedings of the international AAAI conference on web and social media, vol 13, pp 369–380
6. Salminen J, Hopf M, Chowdhury SA et al (2020) Developing an online hate classifier for multiple social media platforms. *Hum Cent Comput Inf Sci* 10:1. <https://doi.org/10.1186/s13673-019-0205-6>
7. Sambaraju R, McVittie C (2020) Examining abuse in online media. *Soc Pers Psychol Compass* 14(3):e12521. <https://doi.org/10.1111/spc3.12521>
8. MacAvaney S, Yao HR, Yang E, Russell K, Goharian N, Frieder O (2019) Hate speech detection: challenges and solutions. *PLoS ONE* 14(8):e0221152. <https://doi.org/10.1371/journal.pone.0221152>
9. Brown A (2017) What is hate speech? Part 1: the myth of hate. *Law Philos* 36(4):419–468. <https://doi.org/10.1007/s10982-017-9297-1>
10. Fortuna P, Nunes S (2018) A survey on automatic detection of hate speech in text. *ACM Comput Surv (CSUR)* 51(4):1–30. <https://doi.org/10.1145/3232676>
11. Srba I, Lenzini G, Pikuliak M, Pecar S (2021) Addressing hate speech with data science: an overview from computer science perspective. *Hate Speech Multidisziplinäre Analysen Und Handlungsoptionen*. https://doi.org/10.1007/978-3-658-31793-5_14
12. Waseem Z, Davidson T, Warmley D, Weber I (2017) Understanding abuse: a typology of abusive language detection subtasks. *arXiv preprint* <http://arxiv.org/abs/1705.09899>
13. Zampieri M, Malmasi S, Nakov P, Rosenthal S, Farra N, Kumar R (2019) Predicting the type and target of offensive posts in social media. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, vol 1, pp 1415–1420. <https://doi.org/10.18653/v1/n19-1144>
14. APA (2013) Diagnostic and statistical manual of mental disorders, 5th edn. American Psychiatric Association, Arlington
15. World Health Organization (2018) International classification of diseases for mortality and morbidity statistics (11th revision). Retrieved from <https://icd.who.int/browse11/l-m/en>
16. Laaksonen S-M, Haapoja J, Kinnunen T, Nelimarkka M, Pöyhtäri R (2020) The datafication of hate: expectations and challenges in automated hate speech monitoring. *Front Big Data*. <https://doi.org/10.3389/fdata.2020.00003>
17. Paz MA, Montero-Díaz J, Moreno-Delgado A (2020) Hate speech: a systematized review. *SAGE Open* 10(4):215824402097302. <https://doi.org/10.1177/2158244020973022>
18. Waqas A, Salminen J, Jung S, Almerekhhi H, Jansen BJ (2019) Mapping online hate: a scientometric analysis on research trends and hotspots in research on online hate. *PLoS ONE* 14(9):e0222194. <https://doi.org/10.1371/journal.pone.0222194>

19. Council of Europe (1997) Recommendation No. R (97) 20 of the Committee of Ministers to member states on “Hate Speech”. Retrieved from <https://rm.coe.int/1680505d5b>
20. EU Commission (2016) Code of conduct on countering illegal hate speech online. Retrieved from <https://ec.europa.eu>
21. UN General Assembly, International Covenant on Civil and Political Rights (1966) United Nations, Treaty Series, vol 999, p 171
22. Brown A (2017) What is hate speech? Part 2: family resemblances. *Law Philos* 36(5):561–613. <https://doi.org/10.1007/s10982-017-9300-x>
23. Calvert C (1997) Hate speech and its harms: a communication theory perspective. *J Commun* 47(1):4–19. <https://doi.org/10.1111/j.1460-2466.1997.tb02690.x>
24. Guterres A (2019) United Nations strategy and plan of action on hate speech. Taken from: https://www.un.org/en/genocideprevention/documents/U_20Strategy. https://doi.org/10.1163/2210-7975_hrd-9915-20190006
25. Parekh B (2012) Is there a case for banning hate speech? In: Herz M, Molnar P (eds) *The content and context of hate speech: rethinking regulation and responses*. Cambridge University Press, Cambridge, pp 37–56. <https://doi.org/10.1017/CBO9781139042871.006>
26. Cohen-Almagor R (2011) Fighting hate and bigotry on the internet. *Policy Int* 3(3):89–114. <https://doi.org/10.2202/1944-2866.1059>
27. Gelber K (2019) Differentiating hate speech: a systemic discrimination approach. *Crit Rev Int Soc Pol Phil*. <https://doi.org/10.1080/013698230.2019.1576006>
28. Bilewicz M, Soral W (2020) Hate speech epidemic. The dynamic effects of derogatory language on intergroup relations and political radicalization. *Polit Psychol* 41(S1):3–33. <https://doi.org/10.1111/pops.12670>
29. Davidson T, Warmsley D, Macy M, Weber I (2017) Automated hate speech detection and the problem of offensive language. arXiv preprint <http://arxiv.org/abs/1703.04009>
30. Community standards, Facebook (2021) Available at https://www.facebook.com/communitystandards/hate_speech/
31. Kalsnes B, Ihlebæk KA (2020) Hiding hate speech: political moderation on Facebook. *Media Cult Soc* 43(2):326–342. <https://doi.org/10.1177/0163443720957562>
32. Švec A, Pikuliak M, Šimko M, Bieliková M (2018) Improving moderation of online discussions via interpretable neural models. In: *Proceedings of the second workshop on abusive language online—ALW ’18 at EMNLP, 2015*, pp 60–65
33. Gillespie T (2018) *Custodians of the internet: platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press, New Haven
34. Chua AYK, Banerjee S (2020) The topic of terrorism on Yahoo! Answers: questions, answers and users’ anonymity. *Aslib J Inf Manag* 72(1):1–16. <https://doi.org/10.1108/AJIM-08-2019-0204>
35. Guimarães SS, Reis JCS, Ribeiro FN, Benevenuto F (2020) Characterizing toxicity on Facebook comments in Brazil. In: *Proceedings of the Brazilian symposium on multimedia and the web (WebMedia ’20)*. Association for Computing Machinery, New York, pp 253–260. <https://doi.org/10.1145/3428658.3430974>
36. Gao L, Huang R (2017) Detecting online hate speech using context aware models. <http://arxiv.org/abs/1710.07395>
37. Wilson RA, Land MK (2021) Hate speech on social media: content moderation in context (September 10, 2020). *Conn Law Rev* 52:1029. Available at SSRN: <https://ssrn.com/abstract=3690616>
38. Tudkuea T, Laeheim K (2014) Development of indicators of cyberbullying among youths in Songkhla Province. *Asian Soc Sci*. <https://doi.org/10.5539/ass.v10n14p74>
39. Molina MD, Sundar SS, Le T, Lee D (2019) “Fake News” is not simply false information: a concept explication and taxonomy of online content. *Am Behav Sci* 65(2):180–212. <https://doi.org/10.1177/0002764219878224>
40. Hudson HK, Fetro JV, Ogletree R (2014) Behavioral indicators and behaviors related to sexting among undergraduate students. *Am J Health Educ* 45(3):183–195. <https://doi.org/10.1080/19325037.2014.901113>
41. Ousidhoum N, Lin Z, Zhang H, Song Y, Yeung D (2019) Multilingual and multi-aspect hate speech analysis. In: *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp 4667–4676. <https://doi.org/10.18653/v1/d19-1474>
42. Waseem Z (2016) Are you a racist or am I seeing things? Annotator influence on hate speech detection on Twitter. In: *Proceedings of the first workshop on NLP and computational social science*, pp 138–142. <https://doi.org/10.18653/v1/w16-5618>
43. 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*. <https://doi.org/10.18653/v1/n16-2013>
44. Azucar D, Marengo D, Settanni M (2018) Predicting the big 5 personality traits from digital footprints on social media: a meta-analysis. *Personal Individ Differ* 124:150–159
45. Bliuc AM, Faulkner N, Jakubowicz A, McGarty C (2018) Online networks of racial hate: a systematic review of 10 years of research on cyber-racism. *Comput Hum Behav* 87:75–86
46. Chaffee SH (1991) *Communication Concepts 1: Explication*. Sage Publications, Newbury Park, CA
47. Krueger RA, Casey MA (2009) *Focus groups: A Practical Guide for Applied Research*. Sage Publications, Thousand Oaks, CA
48. Morgan DL, Krueger RA, Scannell AU (1998) *Planning focus groups*. Sage
49. EU Commission (2020) Countering illegal hate speech online. 5th evaluation of the Code of Conduct. Retrieved from https://ec.europa.eu/info/sites/default/files/codeofconduct_2020_factsheet_12.pdf
50. The UN Refugee Agency (2020) Countering toxic narratives about refugees and migrants. Retrieved from: <https://www.unhcr.org/>
51. CoE-ECRI—Council of Europe—European Commission against Racism and Intolerance: ECRI Report on the Slovak Republic (sixth monitoring cycle); Adopted on 1 October 2020; Published on 8 December 2020, 8 December 2020
52. World Bank (2019) *The little data book on gender 2019*. World Bank, Washington, DC. World Bank. <https://openknowledge.worldbank.org/handle/10986/31689>
53. Robinaugh DJ, Hoekstra RHA, Toner ER, Borsboom D (2019) The network approach to psychopathology: a review of the literature 2008–2018 and an agenda for future research. *Psychol Med* 50(3):353–366. <https://doi.org/10.1017/S0033291719003404>
54. Borsboom D (2017) A network theory of mental disorders. *World Psychiatry* 16(1):5–13. <https://doi.org/10.1002/wps.20375>
55. Pejchal V (2018) Hate speech regulation in post-communist countries: migrant crises in the Czech and Slovak republics. *Int J Crime Justice Soc Democr* 7(2):58. <https://doi.org/10.5204/ijcsd.v7i2.500>
56. Golbeck J, Ashktorab Z, Banjo RO, Berlinger A, Bhagwan S, Buntain C, Cheakalos P, Geller AA, Gergory Q, Gnanasekaran RK, Gunasekaran RR, Hoffman KM, Hottle J, Jienjittlert V, Khare S, Lau R, Martindale MJ, Naik S, Nixon HL, Ramachandran P, Rogers KM, Rogers L, Sarin MS, Shahane G, Thanki J, Vengataraman P, Wan Z, Wu DM (2017) A large labeled corpus for online harassment research. In: *Proceedings of the 2017 ACM on web science conference (WebSci ’17)*. Association for Computing Machinery, New York, pp 229–233. <https://doi.org/10.1145/3091478.3091509>
57. Basile V, Bosco C, Fersini E, Nozza D, Patti V, Pardo FMR, 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 evalu-*

- ation. association for computational linguistics, pp 54–63. <https://doi.org/10.18653/v1/S19-2007>
58. Kumar R, Ojha AK, Malmasi S, Zampieri M (2018). Benchmarking aggression identification in social media. In: Proceedings of the first workshop on trolling, aggression and cyberbullying (TRAC-2018). Association for Computational Linguistics. p 11. <https://aclanthology.org/W18-4401/>
 59. Poletto F, Stranisci M, Sanguinetti M, Patti V, Bosco C (2017) Hate speech annotation: analysis of an Italian Twitter corpus. In: Proceedings of the Fourth Italian conference on computational linguistics (CLiC-it 2017). CEUR-WS, vol 2006. <http://ceur-ws.org/Vol-2006/paper024.pdf>
 60. Sanguinetti M, Poletto F, Bosco C, Patti V, Stranisci M (2018) An Italian Twitter corpus of hate speech against immigrants. In: Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018). European Language Resources Association (ELRA). <https://aclanthology.org/L18-1443>
 61. Chung Y-L, Kuzmenko E, Tekiroglu SS, Guerini M (2019) CONAN—COUNTER NARRATIVES THROUGH NICHESOURCING: a multilingual dataset of responses to fight online hate speech. In: Proceedings of the 57th annual meeting of the association for computational linguistics, pp 2819–2829. Association for Computational Linguistics. <https://doi.org/10.18653/v1/P19-1271>
 62. Bosco C, Felice DO, Poletto F, Sanguinetti M, Maurizio T (2018) Overview of the evalita 2018 hate speech detection task. In: EVALITA 2018—sixth evaluation campaign of natural language processing and speech tools for Italian, vol 2263, pp 1–9. CEUR. <https://iris.unibo.it/retrieve/handle/2318/1686264/465071/paper010.pdf>
 63. Da San Martino G, Barrón-Cedeño A, Wachsmuth H, Petrov R, Nakov P (2020). SemEval-2020 Task 11: detection of propaganda techniques in news articles. In: Proceedings of the fourteenth workshop on semantic evaluation, pp 1377–1414. International Committee for Computational Linguistics. <https://aclanthology.org/2020.semeval-1.186>
 64. Dimitrov D, Ali BB, Shaar S, Alam F, Silvestri F, Firooz H, Nakov P, Da San Martino G (2021) SemEval-2021 task 6: detection of persuasion techniques in texts and images. In: Proceedings of the 15th international workshop on semantic evaluation (SemEval-2021), pp 70–98. <https://doi.org/10.18653/v1/2021.semeval-1.7>
 65. Van Hee C, Lefever E, Hoste V (2018) SemEval-2018 task 3: irony detection in English tweets. In: Proceedings of the 12th international workshop on semantic evaluation. Association for Computational Linguistics, pp 39–50. <https://doi.org/10.18653/v1/S18-1005>
 66. Meaney JA, Wilson SR, Chiruzzo L, Lopez A, Magdy W (2021). SemEval-2021 Task 7: HaHackathon, detecting and rating humor and offense. In: Proceedings of the 15th international workshop on semantic evaluation (SemEval-2021). Association for Computational Linguistics, pp 105–119. <https://aclanthology.org/2021.semeval-1.9.pdf>
 67. Brownstein NC, Louis TA, O'Hagan A, Pendergast J (2019) The role of expert judgment in statistical inference and evidence-based decision-making. *Am Stat* 73(sup1):56–68. <https://doi.org/10.1080/00031305.2018.1529623>
 68. Cassidy MF, Buede D (2009) Does the accuracy of expert judgment comply with common sense: caveat emptor. *Manag Decis* 47(3):454–469. <https://doi.org/10.1108/00251740910946714>
 69. Yarkoni T, Westfall J (2017) Choosing prediction over explanation in psychology: lessons from machine learning. *Perspect Psychol Sci* 12(6):1100–1122. <https://doi.org/10.1177/1745691617693393>
 70. Shrout PE, Fleiss JL (1979) Intraclass correlations: uses in assessing rater reliability. *Psychol Bull* 86(2):420–428. <https://doi.org/10.1037/0033-2909.86.2.420>
 71. OSCE (2020) Annual report 2019. Retrieved from: <http://osr.org>
 72. Fried EI (2020) Lack of theory building and testing impedes progress in the factor and network literature. *Psychol Inq* 31(4):271–288. <https://doi.org/10.1080/1047840x.2020.1853461>
 73. Fersini E, Nozza D, Rosso P (2018a) Overview of the Evalita 2018 task on automatic misogyny identification (AMI). EVALITA evaluation of NLP and speech tools for Italian, vol 12. p 59. <https://boa.unimib.it/retrieve/handle/10281/219426/452274/10281-219426.pdf#page=68>
 74. Fersini E, Rosso P, Anzovino M (2018b) Overview of the task on automatic misogyny identification at IberEval 2018. In: Proceedings of the third workshop on evaluation of human language technologies for Iberian languages (IberEval 2018) co-located with 34th conference of the spanish society for natural language processing (SEPLN 2018). CEUR-WS. <http://ceur-ws.org/Vol-2150/overview-AMI.pdf>
 75. Habernal I, Wachsmuth H, Gurevych I, Stein B (2018). Before name-calling: dynamics and triggers of ad hominem fallacies in web argumentation. In: Proceedings of the 2018 conference of the North American chapter of the Association for Computational Linguistics: human language technologies, volume 1 (long papers), pp 386–396. Association for Computational Linguistics. <https://doi.org/10.18653/v1/N18-1036>
 76. Rosenthal S, Atanasova P, Karadzhov G, Zampieri M, Nakov P (2021) A large-scale semi-supervised dataset for offensive language identification. arXiv preprint <http://arxiv.org/abs/2004.14454>
 77. Constantini G, Epskamp S, Borsboom D, Perugini M, Möttus R, Waldorp LJ, Cramer AO (2015) State of the aRt personality research: a tutorial on network analysis of personality data in R. *J Res Pers* 54:13–29. <https://doi.org/10.1016/j.jrp.2014.07.003>
 78. Robinaugh DJ, Millner AJ, McNally RJ (2016) Identifying highly influential nodes in the complicated grief network. *J Abnorm Psychol* 125(6):747. <https://doi.org/10.1037/abn0000181>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.