Media and Communication (ISSN: 2183–2439)
2021, Volume 9, Issue 1, Pages 171–180
DOI: 10.17645/mac.v9i1.3399

Article

From Insult to Hate Speech: Mapping Offensive Language in German User Comments on Immigration

Sünje Paasch-Colberg *, Christian Strippel, Joachim Trebbe, and Martin Emmer

Institute for Media and Communication Studies, Freie Universität Berlin, 14195 Berlin, Germany;
E-Mails: s.colberg@fu-berlin.de (S.P.-C.), christian.strippel@fu-berlin.de (C.S.), joachim.trebbe@fu-berlin.de (J.T.), martin.emmer@fu-berlin.de (M.E.)

* Corresponding author

Submitted: 26 June 2020 | Accepted: 9 August 2020 | Published: 3 February 2021

Abstract
In recent debates on offensive language in participatory online spaces, the term 'hate speech' has become especially prominent. Originating from a legal context, the term usually refers to violent threats or expressions of prejudice against particular groups on the basis of race, religion, or sexual orientation. However, due to its explicit reference to the emotion of hate, it is also used more colloquially as a general label for any kind of negative expression. This ambiguity leads to misunderstandings in discussions about hate speech and challenges its identification. To meet this challenge, this article provides a modularized framework to differentiate various forms of hate speech and offensive language. On the basis of this framework, we present a text annotation study of 5,031 user comments on the topic of immigration and refuge posted in March 2019 on three German news sites, four Facebook pages, 13 YouTube channels, and one right-wing blog. An in-depth analysis of these comments identifies various types of hate speech and offensive language targeting immigrants and refugees. By exploring typical combinations of labeled attributes, we empirically map the variety of offensive language in the subject area ranging from insults to calls for hate crimes, going beyond the common 'hate/no-hate' dichotomy found in similar studies. The results are discussed with a focus on the grey area between hate speech and offensive language.

Keywords
comment sections; content analysis; Facebook; hate speech; refugees; text annotation; user comments; YouTube

Issue
This article is part of the issue "Dark Participation in Online Communication: The World of the Wicked Web" edited by Thorsten Quandt (University of Münster, Germany).

1. Introduction

In recent years, the use of offensive language in participatory online spaces has increasingly become the subject of public debate and scientific research in many countries (Keipi, Näsi, Oksanen, & Räsänen, 2017). Communication and media scholars analyze this phenomenon using various terms such as 'incivility' (e.g., Coe, Kenski, & Rains, 2014), 'flaming' (e.g., Cho & Kwon, 2015), or 'hate speech' (e.g., Erjavec & Kovačič, 2012). In particular, the term 'hate speech' receives much attention as it has a long tradition in a legal context where it is associated with hate crimes, genocide, and crimes against humanity (Bleich, 2011). In this context, the term refers to violent threats or expressions of prejudice against particular groups on the basis of race, religion, or sexual orientation.

However, due to its explicit reference to the emotion of hate (e.g., Brown, 2017a), 'hate speech' is also understood as a term referring to the expression of hatred (e.g., Post, 2009, p. 123). Accordingly, the term is often used as a general label for various kinds of negative expression by users, including insults and even harsh criticism. This ambiguity leads to fundamental misunderstandings in the discussion about hate speech and challenges its identification, for example, in online user comments (e.g., Davidson, Warmsley, Macy, &

Weber, 2017). Against this background, we formulate the following two research questions: How can we theoretically distinguish hate speech from neighboring concepts (RQ1)? And how can we empirically distinguish various forms of hate speech and offensive language using this theoretical framework (RQ2)? Answering these questions will allow for a more precise measurement of hate speech and offensive language, not only in academic research but also in practical content moderation and community management.

To this end, we introduce a modularized theoretical framework on the basis of which we can operationalize the defining features of hate speech and other forms of offensive language. We will first discuss challenges regarding the definition of hate speech and review how hate speech has been measured in content analyses so far. We then present a new approach to operationalize hate speech for the purpose of content analysis, combining qualitative text annotation and standardized labeling, in which hate speech is not directly identified by coders but rather results from the combination of different characteristics. This approach allows for quantitative description as well as for in-depth analysis of the material. In this article, we focus on the results of a qualitative content analysis of German user comments posted on the topic of immigration and refuge. The in-depth exploration of offensive user comments in the sample shows that our modularized approach allows us to go beyond the common 'hate/no-hate' dichotomy and empirically map the variety of hate speech and offensive language in the subject area.

2. Challenges in Defining Hate Speech

Hate speech is a complex phenomenon and defining it is challenging in several ways. According to Andrew Sellars, "any solution or methodology that purports to present an easy answer to what hate speech is and how it can be dealt with is simply not a product of careful thinking" (Sellars, 2016, p. 5). An important point of disagreement, for example, is the *group reference* of hate speech: On the one hand, some definitions tie the phenomenon to minority groups or list specific group characteristics such as race, religion, gender, or sexual orientation (e.g., Waltman & Mattheis, 2017). On the other hand, some authors stress that hate speech can target every possible group (e.g., Parekh, 2006).

From a theoretical perspective, there are three main approaches in defining hate speech that each *emphasize different aspects*: approaches that (1) refer to the *intentions* behind hate speech; (2) address the *perception* and possible damage of hate speech; and (3) focus on the *content* level and attempt to define hate speech by certain content characteristics (Sellars, 2016, pp. 14–18). For the purpose of content analysis, content-based definitions seem to be most appropriate. For example, Saleem, Dillon, Benesch, and Ruths (2017) focus on speech containing an expression of hatred and use the term 'hateful speech' to emphasize the nuance. Bhikhu Parekh also argues in favor of a content-based understanding and defines 'hate speech' as speech that singles out individuals or groups on the basis of certain characteristics, stigmatizes them and places them outside of society; as such, hate speech "implies hostility, rejection, a wish to harm or destroy, a desire to get the target group out of one's way" (Parekh, 2006, p. 214).

Many scholars approach the heterogeneity of hate speech with rather *broad framework*s: For example, Alexander Brown argues "that the term 'hate speech' is equivocal, that it denotes a family of meanings, for which there is no one overarching precise definition available" (Brown, 2017b, p. 562). He proposes a family resemblances' concept to address hate speech, that is, a "network of similarities overlapping and criss-crossing" (Brown, 2017b, p. 596). Using speech act theory and its basic distinction between locutionary, illocutionary, and perlocutionary speech acts, Sponholz (2017) differentiates hateful speech, hate-fomenting speech, and dangerous speech. The characteristic features of these types are the content, the intent of the speaker, and the context-dependent impact. However, they also differ in respect to language: While hateful speech is typically emotional and uses derogatory language (such as insults or slurs), hate-fomenting speech tends to follow the principles of rationality and reasoning (Sponholz, 2017, pp. 3–5). Nevertheless, empirical studies are challenged with in-between forms of these types, in which the emotional and the rational side of hate speech "coexist to varying degrees" (Keipi et al., 2017, p. 54).

Several authors stress that the emotion or attitude of hatred is not necessarily an essential part of hate speech. Moreover, hate speech can also be rooted, for example, in (religious) beliefs, power relations, boredom, attention-seeking, or negligence (Brown, 2017a). That is why spontaneous and unconsidered forms of hate speech can be expected particularly in participatory online spaces (Brown, 2018, pp. 304–306).

Another problem with hate speech identification is its *overlap with neighboring concepts*. Obviously, hate speech is not the same as dislike or disapproval (Parekh, 2006). However, it is a challenge to consistently identify hate speech and distinguish it from other forms of negative evaluation, since our understanding of hate speech is shaped by changing societal norms, context, and interpretation (Post, 2009; Saleem et al., 2017). This issue becomes evident in low reliability scores for hate speech identification reported in some studies, for example by Ross et al. (2016).

Against this theoretical background, our framework is based on a content-related understanding of hate speech, which seems most appropriate for the purpose of content analysis. In order to avoid assumptions about the intentions of the speaker or possible consequences of a given statement, we argue in favor of the legal origins of the term 'hate speech' and focus on discriminatory content and references to violence within a given statement.

3. Operationalizing Hate Speech for Content Analysis

In order to measure hate speech, the theoretical definitions and dimensions have to be transferred into empirically operable instructions for identifying and categorizing occurrences of the concept. To answer RQ1, in this section we first review existing approaches of measuring hate speech, before we develop our theoretical framework to identify hate speech content in public communication in a multi-dimensional way.

3.1. Existing Approaches of Measuring Hate Speech

Contradicting the elaborate theoretical discussion of hate speech, many empirical studies follow a 'hate/no-hate' dichotomy when categorizing communication content (e.g., Lingiardi et al., 2020). This applies also to non-scientific classification, e.g., in the context of the 'Network Enforcement Law' in Germany, which forces platform companies to identify and block "apparently unlawful" content, including hate speech. Here, it is solely the criterion of 'unlawfulness' that differentiates hate speech from non-hate speech. As a result of this approach, the number of identified (and blocked) items is rather small, given the far-reaching guaranties of free speech in western countries (Facebook, 2020) and does not allow for many insights into the content dimension of hate speech. Thus, many studies in the field avoid a formal law-based, narrow operationalization of hate speech and operate with broader concepts such as 'incivility' (Coe et al., 2014) or 'negative speech' (e.g., Ben-David & Matamoros-Fernández, 2016). A common approach to operationalize such general categories for content analyses is the use of dictionaries that provide pre-categorizations of search terms (e.g., Cho & Kwon, 2015) and that allow for more advanced manual and automated coding of hateful content (e.g., Davidson et al., 2017).

A more differentiated categorization of hate speech can be provided by qualitative approaches (e.g., Ernst et al., 2017), which have the capacity to identify multiple aspects of hate speech and relate them to theoretical dimensions in detail. However, qualitative analyses usually focus on in-depth analysis of specific cases and cannot handle large bodies of text. An example of a multi-dimensional approach to identifying and categorizing different levels of incivility and hate speech on a larger scale following a quantitative approach is presented by Bahador and Kerchner (2019). They applied a computer-aided manual categorization model that ranked the intensity of hate speech on a six-point scale, allowing both for systematic analysis of larger amounts of text and a differentiated recording of aspects of hate speech.

3.2. Introducing a New Approach

Following our theoretical argument, we developed a flexible labeling scheme that measures three key elements of hate speech in text: First, the negative evaluation of a group as a whole, i.e., *negative stereotyping*, is one common element of many hate speech definitions (e.g., Parekh, 2006). For the purpose of our coding scheme, we define negative stereotyping as the attribution of negatively connotated characteristics, roles, or behaviors to the whole group or to individuals on the basis of their group membership (see also, Trebbe, Paasch-Colberg, Greyer, & Fehr, 2017).

Secondly, *dehumanization* is often singled out as one element of hate speech (e.g., Bahador & Kerchner, 2019). On the basis of this literature, we define statements as dehumanization that equate or compare humans with inanimate things (e.g., "scum" or "pack"), animals (e.g., "rats") or inhuman beings (e.g., "demons," "vampires") or characterize humans as savage or animalistic (see also, Maynard & Benesch, 2016). As such, dehumanization is a form of negative stereotyping. However, we agree with Bahador and Kerchner who argue that "dehumanization is a particularly extreme type of negative characterization…and a well-established tool for justifying political violence, and thus merits its own category" (Bahador & Kerchner, 2019, p. 6).

Third, the *expression of violence, harm, or killing* is another important element of hate speech (e.g., Bahador & Kerchner, 2019; Parekh, 2006). Our approach, therefore, defines all statements as hate speech that justify, incite, or threaten physical violence against an individual or a group or that justify, incite, or threaten the killing of individuals or members of a group.

These three elements are measured independently of each other in the sense that they can, but do not have to, apply simultaneously to a comment in order to qualify as hate speech. Thus, our approach allows us to distinguish between forms of hate speech using various combinations of these three elements. In this respect, our approach differs from the scale developed by Bahador and Kerchner (2019), which conceptualizes negative actions, negative characterization, demonizing/dehumanization, violence, and death as different points on a hate speech intensity scale. However, with the help of the hate speech elements of our framework and their various combinations, different types and intensities of hate speech can be identified in the empirical analysis.

For such an analysis, the use of offensive language below the level of hate speech needs to be included, too. Therefore, our coding scheme accounts for three different forms of offensive language that are measured independently of the three hate speech elements: insults and slurs, degrading metaphors, and degrading wordplays.

4. Method

In order to both test this approach empirically and answer our research questions, we conducted a structured text annotation of user comments on news about immigration and refuge to Germany posted in March

2019 in the comment sections of three German news sites (*Compact Magazin*, *Epoch Times*, *Focus Online*), one right-wing blog (*PI news*), four Facebook pages (*FOCUS Online*, *The Epoch Times*, *WELT, Zeit Online*) and 13 YouTube channels (*ARTEde*, *BILD*, *COMPACTTV*, *DW Deutsch*, *Epoch Times Deutsch*, *euronews (deutsch)*, *KenFM*, *Laut Gedacht*, *MrMarxismo*, *Oliver Flesch*, *RT Deutsch*, *tagesschau*, *Tagesschlau*). These sources were selected on the basis of a preliminary study in August 2018, which considered a much broader variety of sources (8 news sites, 3 right-wing blogs, 7 Facebook pages, 31 YouTube channels, and 1 Q&A platform) chosen on the basis of their high reach, their relevance to the public debate of immigration (indicated by the number of user comments), and the variation in their discourse architectures (i.e., comment section, discussion forum, social media, Q&A platform). Following the results of this preliminary study, we selected those sources that contained most hate speech against refugees and immigrants in order to collect as much material as possible for the following analysis. Accordingly, the sample is not designed for a systematic comparison of hate speech in different types of sources.

Using topic related search terms, these sources were screened for articles and posts referring to the topic of immigration and refuge capturing all related user comments. We then randomly selected 178 articles and posts with a total of 6,645 related user comments (for each initial article or post the first up to 50 user comments) for the subsequent analysis. This material was annotated using the *BRAT rapid annotation tool*, a browser-based software for structured text annotation (Stenetorp et al., 2012).

The method of structured text annotation includes that each text is examined for relevant words, sentences, or sections ('entities'), which are then selected and labeled with predefined categories ('entity attributes'). Thus, this method is basically a combination of the inductive identification of relevant text segments as we know it from computer-assisted qualitative text analysis and the assignment of codes to these text segments as we know it from standardized content analysis. As such, it is particularly helpful for content analysis as the classification is explicitly related to specific parts of a text, which are at the same time recorded for subsequent analysis. This allows us to conduct both a standardized and a qualitative content analysis of the annotated user comments.

Both the methodological approach and our focus on immigration and refuge to Germany were chosen due to the broader research context of this study, which aims at the automatization of detecting hate speech against refugees and immigrants in German user comments. For this reason, we will first take a closer look at the situation in Germany (see Section 4.1). While our methodological approach is suitable for analyzing hate speech against various groups (see Section 4.2), the results presented in this article are limited to hate speech against refugees and immigrants. This is not to say that only refugees and immigrants are recently affected by hate speech in Germany; anti-Semitic hate speech, for example, has dramatically increased again as well (Hänel, 2020; Schwarz-Friesel, 2019). Nevertheless, we consider a focus on a specific target group to be helpful in order to distinguish hate speech from neighboring concepts.

4.1. Immigration and Refuge to Germany

The topic of immigration and refuge was chosen for our case study as it has been heavily discussed in public since 2015, when the German Chancellor Angela Merkel decided to keep the state's borders open and the number of refugees entering Germany rose sharply. Even though the number of asylum applications dropped drastically in 2017 and continues to decrease (Bundesamt für Migration und Flüchtlinge, 2020), questions of immigration have repeatedly triggered heated political debates in Germany in the following years and have long been high on the media agenda (e.g., Krüger & Zapf-Schramm, 2019). The public opinion was increasingly divided on the issue, and dissatisfaction with political institutions and the processes that deal with it is widespread (Arlt, Schumann, & Wolling, 2020).

This social division has become apparent, for example, in anti-immigration protests (Bennhold, 2018), the rise of the populist extreme right-wing party 'Alternative für Deutschland' (Bennhold, 2018) and a growing mistrust regarding the accuracy of media coverage on refugees (Arlt & Wolling, 2016). However, this issue is of particular relevance as the growing online hate speech against refugees and immigrants has been accompanied by an increase in racist hate crimes against these groups in Germany in recent years (Eddy, 2020; Hille, 2020). Examples include the attacks in Hanau (February 2020), Halle (October 2019) and Munich (June 2016) as well as the murder of the Hessian politician Walter Lübcke, who publicly supported liberal refugee politics (June 2019). There is reason to believe that such hate crimes are verbally prepared, socially backed, and ideologically legitimized by hate speech on far-right websites and forums, but also on social media and in comment sections of news websites (see e.g., Scholz, 2020).

4.2. Annotation Rules and Coding Scheme

On the basis of a detailed theory-based manual, three trained coders annotated the user comments in our sample following three steps: First, the initial article or post was read and checked for thematic relevance. The net sample contains 135 relevant articles or posts with 5,031 corresponding user comments.

In the second step, all judgments of individuals or groups within these comments were identified and annotated as 'entities' on a sentence level. Thereby, a judgment is defined as a statement expressing an opinion or an evaluation of the person/group by ascribing negative characteristics, roles, or behavior to it. Such judgments

can be recognized by attributions of adjectives, judgmental subjectivizations, the attribution of behavior that violates social standards, or by association with certain consequences (e.g., damage). In addition to such explicit judgments, the coders were also instructed to identify and annotate implicit judgments, expressed by rhetorical questions, ironic statements, or historical references. In order to capture such implicit forms as validly as possible, the manual includes dimensions and examples taken from a qualitative expert survey of German community managers (Paasch-Colberg, Strippel, Laugwitz, Emmer, & Trebbe, 2020), qualitative pre-coding and literature.

In the third step, all annotated judgments were further qualified by attributing predefined labels to them, such as the targets of judgment (e.g., politicians, journalists/media, German citizens, right-wing groups, Muslims, and refugees/immigrants) and the subject of judgment (e.g., culture, sexuality or character/behavior). Judgments that were attached to a specific group membership (i.e., ethnicity, nationality, religion, gender, profession) and are thus stereotyping were labeled accordingly. It was further specified whether a judgment includes a dehumanization (as defined in Section 3.2) or a response to the target group. Possible responses range from non-violent forms (i.e., rejection) to violent forms (i.e., the legitimization, threat or call for physical violence or killing of the person/group).

Finally, the manual contains three attributes to specify different forms of offensive language, i.e., insults and slurs, derogatory metaphors and comparisons as well as derogatory wordplays. The manual includes examples for these forms of offensive language used in German user comments, which were drawn primarily from the aforementioned expert survey.

The context unit of the annotation of judgments in a user comment were the news item or social media posting and the preceding comments, i.e., the coders were instructed to use textual references within this context to identify judgments.

4.3. Data Analysis

A qualitative content analysis was conducted for all user comments in the sample that contains at least one element of hate speech or offensive language according to our framework. The hate speech and offensive language elements described in Section 3.2 were thus used as deductive categories to pre-structure the material. In the first step, the comments in these pre-set categories were close-read in order to describe and exemplify the categories as such. In the second step, the analysis was focused on those comments that target refugees or other immigrants and segmented into (1) comments that qualify as hate speech and (2) comments that qualify as offensive language but not as hate speech. These comments were then further explored using the technique of structuring qualitative content analysis according to Mayring (2015). This form of qualitative content analysis focuses on patterns and co-occurrences of selected characteristics in the material and aims at the description of different types in the material (Mayring, 2015, pp. 103–106; Schreier, 2014). To assure consistency, the material was close-read by two researchers independently and inconsistencies were resolved in discussing.

5. Results

In our sample of 5,031 user comments, 2,602 negative judgments were identified. Hate speech was identified in 25% of the judgments (n = 701) and, since a comment can contain more than one judgment, in 11% of the comments (n = 538). With regard to the three hate speech elements, negative stereotyping is by far the most frequent element. Every fifth judgment in our sample (n = 539) uses negative stereotypes while only 155 judgments (6%) dehumanize the target. Calls for violence or death were identified even less frequently (n = 56 and n = 57). The majority of judgments with hate speech are targeting the group of refugees and immigrants.

Offensive language is more frequent in our sample than hate speech. And if offensive language is used in a comment, it is often used more than once: Offensive language was identified in 16% of the comments (n = 796) and 38% of the judgments (n = 1,070). About 60% of these judgments use offensive language without qualifying as hate speech according to our framework.

5.1. Describing Hate Speech in German User Comments on Immigration and Refuge

The following sections present examples of user comments that contain potentially offensive and upsetting terms, particularly racist and islamophobic. They are solely used as examples to illustrate the results of this research and do not reflect the views of the authors in any way. The user comments were translated, the German originals can be found in the supplementary document.

In a first step, we close-read the user comments to illustrate in more depth how the three hate speech elements defined in Section 3.2 can be identified. This aims at describing the main categories of our framework, illustrates them with examples and thus makes them applicable for further analysis in the field. As Table 1 shows, hate speech in our sample is expressed through different kinds of rhetoric and can be identified by different indicators. At the least extreme level, groups are negatively stereotyped by referring to the average or majority of its members or by calling a behavior or role typical for the group. Another form of stereotyping is to criticize the behavior of a group as a negative deviation from supposedly normal behavior.

Dehumanizing hate speech refers to humans as things, animals, or other inhuman beings, considered inferior, disgusting, or dangerous.

Table 1. Description of hate speech elements.

Hate speech element	Example (English translation)
Negative stereotyping	
Referring to everybody, most people, or the average or typical person	"These newcomers are all potential killers, they pull out their knives on every little thing"
Social groups, religious groups, professional roles, or nationalities are generalized	"Muslim and Black African, the recipe for murder and manslaughter"
Critique is tied to the deviation of 'normality'	"Nowhere else in the world do criminal asylum seekers get so much support and so many murderers can run free like they do here"
Dehumanization	
Humans are equated as or compared to inanimate things	"Whoever takes the stuff out has to be well paid. What kind of sewer man digs in shit without proper pay?"
Humans are equated as or compared to animals or inhuman beings	"Unfortunately, the money is not enough to get rid of even a small portion of these parasites"
Violence and killing	
Fantasies of violence/killing	"Let the cops beat him until he's crippled! Then fly him across the desert and throw him out"
Violence/killing as only effective means or remedy	"The only thing that helps is violence"
Violence/killing as a right/appropriate solution	"It would have been faster, cheaper and more sustainable to just shoot him"
Specific calls for violence/killing	"When all subjects are still in deep sleep, let's blow up the asylum center!"

The user comments in the category 'violence and killing' address a broad spectrum of violence, ranging from general physical violence and more specific forms such as sexual violence, violence in law enforcement, extreme punishment (i. e., forced labor, torture), or (civil) war to murder, suicide, deadly revenge, or death penalty. Furthermore, the category includes violent fantasies, rhetoric describing violence or killing as the only effective means or the appropriate solution, and specific calls for violent action or killing.

5.2. Mapping the Variety of Hate Speech and Offensive Language towards Immigrants and Refugees

To answer RQ2, we then used our multi-dimensional annotations to identify patterns by grouping the user comments in our sample to general types. To derive the types, the common occurrence of the labeled characteristics (including the three hate speech elements and forms of offensive language as defined in Section 3.2) was examined. For those user comments that target immigrants or refugees, five types of hate speech emerged which partly build on each other, so that their borders tend to be blurry; also, individual user comments may recur to more than one type at once.

Racist othering: Key characteristics of this type are an 'us against them'-rhetoric and a sharp devaluation of the designated out-group. At least implicitly, this type is the basic motive of hate speech. The element of negative stereotyping applies to all comments of this type: Immigrants and refugees are negatively stereotyped (e.g., as lazy, stupid, rude), and framed as a burden and imposition to the ingroup, as this example shows: "Anyone who comes here to participate in what our forefathers built, and their ancestors did not contribute anything at all, is unwanted because he is only scrounging, no matter what else he says or does." The devaluation is often associated with descriptions of an allegedly abnormal sexual life, as in this example: "Illiterate Afros, Arabs, and Afghanis have no access to women of their ethnicity and consequently suffer a hormonal emergency."

Racist criminalization: This type is a special form of negative stereotyping, which focuses on the description of immigrants and refugees as a threat. Crime is culturized and associated particularly with the male gender. In this context, it is striking that the knife is coined as the central tool of crime, shaping the image of an uncivilized wild: "We live in hard times in which one must constantly count on getting a knife from foreigners, who were raised differently." Forms of self-victimization are also identified, whereby the sexual motif reappears as the narrative of the threatened German woman: "These murderers, rapists, and thieves from Morocco, Algeria, or Mauritania cause the most damage to the population and are therefore Merkel's darlings."

Dehumanization: This type builds on the previous types, but is characterized by an additional dehumanization of the target group; in other words, comments of this type are distinguished by the common presence of the elements of negative stereotyping and dehumaniza-

tion. Immigrants and refugees are compared or referred to as non-human things or beings that are connoted as inferior, disgusting or even dangerous, as this example shows: "The scum from which the whole world protects itself is integrated into the social systems here." The second example is a hybrid of racist criminalization, expressed through a play on the words criminal and migrant, and dehumanization: "These Crimigrants are predators. They lurk and choose their victims."

Raging hate: User comments of this type are distinguished by the element of violence and killing. Physical violence against immigrants and refugees or even their death is legitimized or demanded; other comments imply fantasies of violence and killing as acts of revenge. Some comments of this type also contain the element of dehumanization, as if to justify (lethal) violence. Further, this type is characterized by the use of offensive language, i. e., insults, which imply malice, cynicism, disgust, and aggression: "That filthy mutt is still alive?"

Call for hate crimes: The main characteristic of this type is the occurrence of the element of violence and killing. However, in contrast to the type of raging hate, this is done without the use of offensive language, but in a distanced and calm form, as this example shows: "The attacker should be shot, stabbed, or beaten to death immediately. Another language is not understood by Muslim Africans. Otherwise they understand: Keep up the good work." Calls for hate crimes often use negative stereotyping (e. g., by criminalizing immigrants and refugees) and dehumanization as a justifying rhetoric: "They should not be stoned, but fed to the lions. Something like that must not live."

We further analyzed the use of offensive language in the user comments, to assess its role for the five types of hate speech as well as the grey area that exists in the demarcation of hate speech and offensive language. The in-depth analysis showed that comments targeting immigrants and refugees use different forms of offensive language.

First, the target group is described with common racial slurs and insults. User comments that contain racial insults but none of the hate speech elements described in Section 3.2 do not qualify as hate speech according to our framework. However, they would do so on the basis of other definitions in the literature (e.g., Saleem et al., 2017). Thus, they are clearly sitting in a grey area between hate speech and offensive language.

In addition, derogatory group labels are identified that either use neologisms or wordplays. The distinction between this form and common racial insults is temporary and fluent, as such, these labels can also be considered as a grey area. However, they are difficult to capture in standardized approaches and require special knowledge. The same holds for ironic group labels (e.g., "gold pieces") that are highly context-sensitive.

Another form of offensive language can be referred to as distancing, as it denies refugees their legal status, e.g., by using quotation marks ("so-called 'refugees'"), adjectives such as "alleged" or neologisms such as "refugee actors." Distancing can be understood as a preliminary stage to racist othering. Finally, user comments referring to immigrants and refugees also use common insults (e.g., "wanker") against them without referring to the group of refugees as a whole. Therefore, this form qualifies as incivility (in the sense of impoliteness) but clearly not as hate speech.

Offensive language was found to be used in all hate speech types, however, the type 'call for hate crimes' seems to be an exception to that.

6. Conclusions

In this article we developed a new approach to hate speech definition and identification that aims at solving some of the described challenges in the field of research and goes beyond the common 'hate/no-hate' dichotomy. To add more depth to the concept of hate speech and answering RQ1, our theoretical approach first developed a multi-dimensional understanding of the term based on the dimensions of discriminatory content and references to violence, which in the second step was measured using a set of independent labels. In contrast to most existing studies in the field, hate speech thus could be measured indirectly and in a multi-dimensional way.

In a structuring content analysis of user comments targeting immigrants and refugees, we showed how this approach allows an in-depth analysis of the character of hate speech statements in a content analysis as well as, in a second step, the development of distinct types of hate speech that form a dark spectrum of discrimination and violence-related statements. Answering RQ2, our approach captures recurring patterns of hate speech, as identified and described in other qualitative studies, and enables their standardized measurement: The types of racist othering, racist criminalization, and dehumanization correspond largely to some of the hate myths identified by Waltman and Mattheis (2017) in hate novels of US-American white supremacists. Dehumanization and racist criminalization resemble closely some of the justificatory hate speech mechanisms identified by Maynard and Benesch (2016, pp. 80–82) in the context of mass atrocities.

The results further show that two of the hate speech types are characterized by a special relationship to language and thus deepens our knowledge on the role of offensive language for hate speech: While the use of offensive language is constitutive for 'raging hate,' the type 'call for hate crimes' is characterized by a quite rational language. Hence, the empirical analysis supports our argument that a deeper theoretical conceptualization of hate speech and offensive language as two distinct dimensions allows for much more detailed insights into the nature of this phenomenon.

Our case study is limited in several ways. Firstly, our analysis addresses hate speech in user comments. While this is a relevant perspective because most hate content

emerges in this sphere, it is only one facet of the problem of offensive language in participatory online discussions. In order to better understand the dynamics of escalating discussions, future studies should therefore consider the broader context and, for example, analyze discriminatory speech in the discussed news pieces and social media posts themselves.

Secondly, the analysis is not based on a representative set of sources, but biased by the right-wing news sites and the right-wing blog selected for analysis. Therefore, our typology can only be preliminary and must be validated and quantified in further studies. Such further empirical applications of our framework should in particular consider the differences between different types of sources systematically.

Thirdly, we limited our analysis to hate speech targeting immigrants and refugees, as this seems to be particularly relevant against the background of recent hate crimes in Germany (see Section 4.1). Nevertheless, the question of what forms of hate speech are used to target other social groups should definitely be answered in future studies.

Finally, capturing implicit forms of hate speech is quite difficult. In order to prevent corresponding user comments from being deleted directly, hate speech is sometimes strategically disguised (e.g., Warner & Hirschberg, 2012). Another challenge with regard to right-wing blogs and websites in specific is the strategy of right-wing extremists to use their websites for image control and to avoid open racisms and calls for violence (e.g., Gerstenfeld, Grant, & Chiang, 2003). Through a previous expert survey, we were able to supplement our manual with many current examples of implicit hate speech. However, this form of hate speech can change significantly over time, which is why our manual at this point is more of a snapshot that needs updating for and through future research. Moreover, our framework focuses on text and does not include forms of hate speech expressed by non-textual communication, such as memes for example.

Nevertheless, we argue that our framework provides a sensitive tool to describe the prevalence of hate speech in more detail than existing approaches, while also considering borderline cases and rhetoric that prepare hate speech. This extended perspective on the phenomenon of hate speech is promising to better understand escalating dynamics in participatory online spaces and to empirically test different counter-measures, for example. This is of particular importance for practical social media community and content management. When integrated into existing (semi-)automated content management systems, such a tool that distinguishes between several types and intensities of incivility and hate speech may contribute to more adequate strategies of dealing with disturbing content than many of the existing keyword-based and binary 'hate/no-hate' systems. This is even more important as simple deletion of 'hate'-labeled postings often raises concerns of censorship, particularly when measurement is blurry and mistakenly covers also non-hate speech content.

Finally, with reference to the various hate speech definitions in the literature, we want to point out the flexibility of our approach: It can be adapted to answer specific research questions and make different or broader hate speech definitions operational for content analysis, e.g., definitions of 'hateful speech' that would include racial insults but exclude the element of negative stereotyping. By combining it with surveys or experiments, the content-related perspective of our approach can also be related to other perspectives on hate speech in order to provide additional insights, for example, into the interplay of text characteristics and their perception by different population groups.

Acknowledgments

This research is part of the project "NOHATE—Overcoming crises in public communication about refugees, migration, foreigners," funded by the German Federal Ministry of Education and Research (grant number 01UG1735AX). The authors would like to thank Laura Laugwitz for her support of the study and the coding students for their work.

Conflict of Interests

The authors declare no conflict of interests.

Supplementary Material

Supplementary material for this article is available online in the format provided by the author (unedited).

References

Arlt, D., Schumann, C., & Wolling, J. (2020). Upset with the refugee policy: Exploring the relations between policy malaise, media use, trust in news media, and issue fatigue. *Communications*. Advance online publication. https://doi.org/10.1515/commun-2019-0110

Arlt, D., & Wolling, J. (2016). The refugees: Threatening or beneficial? Exploring the effects of positive and negative attitudes and communication on hostile media perceptions. *Global Media Journal German Edition*, *6*(1), 1–21.

Bahador, B., & Kerchner, D. (2019). *Monitoring hate speech in the US media* (Working Paper). Washington, DC: The George Washington University.

Ben-David, A., & Matamoros-Fernández, A. (2016). Hate speech and covert discrimination on social media: Monitoring the Facebook pages of extreme-right political parties in Spain. *International Journal of Communication*, *10*, 1167–1193.

Bennhold, K. (2018, August 13). Chemnitz protests show new strength of Germany's far right. *The New*

York Times. Retrieved from https://www.nytimes.com/2018/08/30/world/europe/germany-neo-nazi-protests-chemnitz.html

Bleich, E. (2011). The rise of hate speech and hate crime laws in liberal democracies. *Journal of Ethnic and Migration Studies*, *37*(6), 917–934.

Brown, A. (2017a). What is hate speech? Part 1: The myth of hate. *Law and Philosophy*, *36*(4), 419–468.

Brown, A. (2017b). What is hate speech? Part 2: Family resemblances. *Law and Philosophy*, *36*(5), 561–613.

Brown, A. (2018). What is so special about online (as compared to offline) hate speech? *Ethnicities*, *18*(3), 297–326.

Bundesamt für Migration und Flüchtlinge. (2020). *Asyl und Flüchtlingsschutz. Aktuelle Zahlen* [Asylum and refugee protection. Current figures] (04/2020). Berlin: Bundesamt für Migration und Flüchtlinge.

Cho, D., & Kwon, K. H. (2015). The impacts of identity verification and disclosure of social cues on flaming in online user comments. *Computers in Human Behavior*, *51*, 363–372.

Coe, K., Kenski, K., & Rains, S. A. (2014). Online and uncivil? Patterns and determinants of incivility in newspaper website comments. *Journal of Communication*, *64*(4), 658–679.

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). *Automated hate speech detection and the problem of offensive language.* Paper presented at the Eleventh International Conference on Web and Social Media, Montreal, Canada.

Eddy, M. (2020, February 21). Far-right terrorism is no. 1 threat, Germany is told after attack. *The New York Times*. Retrieved from https://www.nytimes.com/2020/02/21/world/europe/germany-shooting-terrorism.html

Erjavec, K., & Kovačič, M. P. (2012). "You don't understand, this is a new war!" Analysis of hate speech in news web sites' comments. *Mass Communication and Society*, *15*(6), 899–920.

Ernst, J., Schmitt, J. B., Rieger, D., Beier, A. K., Vorderer, P., Bente, G., & Roth, H.-J. (2017). Hate beneath the counter speech? A qualitative content analysis of user comments on YouTube related to counter speech videos. *Journal for Deradicalization*, *2017*(10), 1–49.

Facebook. (2020). *NetzDG Transparenzbericht* [Network enforcement act: Transparency report]. Menlo Park, CA: Facebook. Retrieved from https://about.fb.com/wp-content/uploads/2020/01/facebook_netzdg_Januar_2020_German.pdf

Gerstenfeld, P. B., Grant, D. R., & Chiang, C.-P. (2003). Hate online: A content analysis of extremist Internet sites. *Analyses of Social Issues and Public Policy*, *3*(1), 29–44.

Hänel, L. (2020, May 7). Germany: Anti-semitism despite remembrance culture. *Deutsche Welle*. Retrieved from https://www.dw.com/en/germany-anti-semitism-despite-remembrance-culture/a-53360634

Hille, P. (2020, February 20). Right-wing terror in Germany: A timeline. *Deutsche Welle*. Retrieved from https://www.dw.com/en/right-wing-terror-in-germany-a-timeline/a-52451976

Keipi, T., Näsi, M., Oksanen, A., & Räsänen, P. (2017). *Online hate and harmful content: Cross-national perspectives*. London: Routledge.

Krüger, U. M., & Zapf-Schramm, T. (2019). InfoMonitor 2018: GroKo und Migrationsdebatte prägen die Fernsehnachrichten. Analyse der Nachrichtensendungen von Das Erste, ZDF, RTL und Sat.1 [InfoMonitor 2018: Grand coalition and migration debate dominate the television news—Analysis of the news programs of Das Erste, ZDF, RTL and Sat.1]. *Media Perspektiven*, *2019*(2), 44–73.

Maynard, J., & Benesch, S. (2016). Dangerous speech and dangerous ideology: An integrated model for monitoring and prevention. *Genocide Studies and Prevention*, *9*(3), 70–95.

Lingiardi, V., Carone, N., Semeraro, G., Musto, C., D'Amico, M., & Brena, S. (2020). Mapping twitter hate speech towards social and sexual minorities: A lexicon-based approach to semantic content analysis. *Behaviour & InformationTechnology*, *39*(7), 711–721.

Mayring, P. (2015). *Qualitative Inhaltsanalyse. Grundlagen und Techniken* [Qualitative content analysis: Basics and techniques] (12th ed.). Weinheim and Basel: Beltz.

Paasch-Colberg, S., Strippel, C., Laugwitz, L., Emmer, M., & Trebbe, J. (2020). Moderationsfaktoren: Ein Ansatz zur Analyse von Selektionsentscheidungen im Community Management [Moderation factors: A framework to the analysis of selection decisions in community management]. In V. Gehrau, A. Waldherr, & A. Scholl (Eds.), *Integration durch Kommunikation. Jahrbuch der Publizistik- und Kommunikationswissenschaft 2019* [Integration through communication. Yearbook of Journalism and Communication Studies 2019] (pp. 109–119). Muenster: DGPuK.

Parekh, B. (2006). Hate speech. Is there a case for banning? *Public Policy Research*, *12*(4), 213–223.

Post, R. (2009). Hate speech. In I. Hare & J. Weinstein (Eds.), *Extreme speech and democracy* (pp. 123–138). Oxford: Oxford University Press.

Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., & Wojatzki, M. (2016). *Measuring the reliability of hate speech annotations: The case of the European refugee crisis.* Paper presented at the 3rd Workshop on Natural Language Processing for Computer-Mediated Communication, Bochum, Germany.

Saleem, H. M., Dillon, K. P., Benesch, S., & Ruths, D. (2017). *A web of hate: Tackling hateful speech in online social spaces*. Paper presented at the first Workshop on Text Analytics for Cybersecurity and Online Safety (TA-COS), Portorož, Slovenia.

Scholz, K.-A. (2020, February 21). How the inter-

net fosters far-right radicalization. *Deutsche Welle*. Retrieved from https://www.dw.com/en/how-the-internet-fosters-far-right-radicalization/a-52471852

Schreier, M. (2014). Qualitative content analysis. In U. Flick (Ed.), *The Sage handbook of qualitative data analysis* (pp. 1–19). London: Sage.

Schwarz-Friesel, M. (2019). "Antisemitism 2.0": The spreading of Jew-hatred on the World Wide Web. In A. Lange, K. Mayerhofer, D. Porat, & L. H. Schiffman (Eds.), *Comprehending and confronting antisemitism: A multi-faceted approach* (pp. 311–338). Boston, MA: De Gruyter.

Sellars, A. F. (2016). *Defining hate speech* (Research publication No. 2016–20). Cambridge, MA: Berkman Klein Center.

Sponholz, L. (2017). *Tackling hate speech with counter speech? Practices of contradiction ans their effects*. Paper presented at the International Conference Worlds of Contradiction, Bremen, Germany.

Stenetorp, P., Pyysalo, S., Topic, G., Ohta, T., Ananiadou, S., & Tsujii, J. I. (2012). *BRAT: A web-based tool for NLP-assisted text annotation*. Paper presented at the 13th Conference of the European Chapter of the Association for Computational Linguistics, Avignon, France.

Trebbe, J., Paasch-Colberg, S., Greyer, J., & Fehr, A. (2017). Media representation: Racial and ethnic stereotypes. In P. Rössler (Ed.), *The international encyclopedia of media effects*. Hoboken, NJ: John Wiley & Sons.

Waltman, M. S., & Mattheis, A. A. (2017). Understanding hate speech. In *Oxford Research Encyclopedia of Communication*. New York, NY: Oxford University Press. Retrieved from http://communication.oxfordre.com/view/10.1093/acrefore/9780190228613.001.0001/acrefore-9780190228613-e-422

Warner, W., & Hirschberg, J. (2012). Detecting hate speech on the world wide web. In S. Owsley Sood, M. Nagarajan, & M. Gamon (Eds.), *Proceedings of the 2012 Workshop on Language in Social Media* (pp. 19–26). Montreal: Association for Computational Linguistics.

About the Authors

Sünje Paasch-Colberg is a Postdoc Researcher at the Institute for Media and Communication Studies at Freie Universität Berlin, Germany. She currently works in the research project "NOHATE—Overcoming crises in public communication about refugees, migration, foreigners." Her research interests are digital communication, migration, integration and media, and content analysis. In her dissertation she worked on media effects in election campaigns.

Christian Strippel is a Research Assistant and PhD candidate at the Institute for Media and Communication Studies at Freie Universität Berlin, Germany. He currently works in the research project "NOHATE—Overcoming crises in public communication about refugees, migration, foreigners." His research interests include digital communication, media use, public sphere theory, and sociology of scientific knowledge.

Joachim Trebbe is Professor for media analysis and research methods at the Institute for Media and Communication Studies at Freie Universität Berlin, Germany. His research interests include the media use of migrants in Germany, the media representations of ethnic minorities, and research methods for television content analyses.

Martin Emmer is Professor for media use research at the Institute for Media and Communication Studies at Freie Universität Berlin and Principle Investigator of the research group "Digital Citizenship" at the Weizenbaum Institute, Berlin, Germany. His research focusses on digital communication and digital publics, political communication and participation, and methods of empirical communication research.