Hate Speech Criteria: A Modular Approach to Task-Specific Hate Speech Definitions

Urja Khurana[1*] and **Ivar Vermeulen**[2] and **Eric Nalisnick**[3]
and **Marloes van Noorloos**[4] and **Antske Fokkens**[1,5*]
[1]Computational Linguistics and Text Mining Lab, Vrije Universiteit Amsterdam
[2]Department of Communication Science, Vrije Universiteit Amsterdam
[3]Informatics Institute, University of Amsterdam
[4]Department of Criminal Law, Tilburg University
[5]Dept. of Mathematics and Computerscience, Eindhoven University of Technology

Abstract

Offensive Content Warning: This paper contains offensive language only for providing examples that clarify this research and do not reflect the authors' opinions. Please be aware that these examples are offensive and may cause you distress.

The subjectivity of recognizing *hate speech* makes it a complex task. This is also reflected by different and incomplete definitions in NLP. We present *hate speech* criteria, developed with perspectives from law and social science, with the aim of helping researchers create more precise definitions and annotation guidelines on five aspects: (1) target groups, (2) dominance, (3) perpetrator characteristics, (4) type of negative group reference, and the (5) type of potential consequences/effects. Definitions can be structured so that they cover a more broad or more narrow phenomenon. As such, conscious choices can be made on specifying criteria or leaving them open. We argue that the goal and exact task developers have in mind should determine how the scope of *hate speech* is defined. We provide an overview of the properties of English datasets from `hatespeechdata.com` that may help select the most suitable dataset for a specific scenario.

1 Introduction

The surge in online *hate speech* has resulted in an increased need for its automatic detection. Its presence can be highly consequential as it creates an unsafe environment and threatens the freedom of speech (Kiritchenko et al., 2021). Effects of hate speech range from a personal level (e.g. anxiety or stress (Cervone et al., 2021)) to societal level (e.g. discrimination or violence (Waldron, 2012)) and such speech can disrupt social debate severely (Vidgen and Derczynski, 2020). Due to the large volumes of data on social media, automatizing the task is essential as hate speech can violate the law, depending on the country, in addition to its negative consequences in society. This makes automatic hate speech detection a very important task that needs to be carried out responsibly.

What is considered *hate speech* is subjective (Fortuna et al., 2020), there are a variety of valid viewpoints on what does (not) fall under this concept. Current hate speech datasets in NLP reflect this, having similar yet (subtly) different or incomplete definitions. For instance, similar terms are used interchangeably across publications and datasets, e.g. abusive, offensive, or toxic (Madukwe et al., 2020; Fortuna et al., 2020). We posit that a clear relation to (membership of) a target group of the victim sets *hate speech* apart from other forms of toxic or abusive language. Underspecified definitions and guidelines increase the level of subjectivity in annotations. This subjectivity propagates into the model, which can lead to biased models (Sap et al., 2019; Davidson et al., 2019). Even if annotations are systematic, it may remain unclear which phenomena (e.g. target groups or types of abusive) are covered and thus captured by models.

It depends on the task for which a dataset is created whether subjectivity is desired or not. We will argue that, even for scenarios where the goal is to collect multiple viewpoints, it is important to clearly define on what aspects of the phenomenon this subjectivity is sought. At the same time, we must keep in mind that it is impossible to fully remove subjectivity when determining whether something is *hate speech* or not. Even in law, where *hate speech* is aimed to be defined as objectively as possible, it is inevitable that courts have difficulties interpreting such a context-dependent and sensitive topic in consistent and predictable ways (Van Noorloos, 2014b). Nevertheless, a clear definition can help reduce subjectivity to borderline cases.

What would then be a good definition of *hate*

*Main correspondence: `u.khurana@vu.nl` and `antske.fokkens@vu.nl`

Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH), pages 176 - 191
July 14, 2022 ©2022 Association for Computational Linguistics

speech? We advocate that a good definition of *hate speech* starts with a good understanding of the intention that it serves. For instance, a social media platform may want a *broader* consideration of *hate speech*, as it needs to keep its platform safe, in comparison to a law-enforcing model, which has to take legal action only when there is a clear punishable presence. One may decide to focus on hate speech towards one specific group or to keep this completely open to investigate which groups are considered potential targets by (crowd) annotators.

Rather than specifying what "the" definition should be, we provide a meta-prescriptive setup to construct definitions and guidelines through a modular approach, where modifications can be made according to the task at hand. Concretely, we propose five criteria which should be taken into account when defining *hate speech*[1] and creating annotation guidelines: (1) target groups, (2) social status of target groups, (3) perpetrator, (4) type of negative reference, (5) type of potential effect/consequence. These criteria were developed with insights from law and social science. We provide an overview of all English datasets from *www.hatespeechdata.com* according to these criteria, so that people working with a specific definition in mind can easily identify which existing English datasets[2] may be of direct use or provide a good starting point.

2 Background and Motivation

In this section, we describe related work that provided the motivation and insights for the operationalization we propose in this paper.

2.1 Annotations for Hate Speech Tasks

The quality of annotations directly influences the quality of hate speech detection models trained on the annotated data. The subjective nature of the task makes obtaining high inter-annotator agreement, often used as a quality metric for annotations, difficult (Talat et al., 2017). Awal et al. (2020) analyze and find evidence for inconsistency in the annotation for different widely-used hate speech datasets. They discover that some of the retweets in the dataset of Founta et al. (2018) have different labels while the tweet is the same, as also found by Isaksen and Gambäck (2020).

Demographic factors such as language, age, and educational background have an impact on how annotation is done (Al Kuwatly et al., 2020; Schmidt and Wiegand, 2017) as well as expertise (Talat, 2016). This subjectivity of the annotator also brings in possible biases of their own, as illustrated in Sap et al. (2019), Davidson et al. (2019), and Talat (2016). Sap et al. (2019) show that priming annotators and making them aware of their racial bias can decrease such inclinations which could stem from misunderstanding the intent of the text.

Vidgen and Derczynski (2020) point out that annotators need more appropriate guidelines with clear examples to get better annotations. They also argue that it is good practice to create training sets in such a way that they address the task. We follow this point of view and start from the assumption that any annotation task (whether it is intended for training, evaluation or exploration) starts with establishing the purpose of the annotations: How will the system trained on them be used? What is investigated in case of exploratory research?

2.2 Task Specific Annotations

One factor in settling how annotations can support the purpose of the task is how much subjectivity is desired. Röttger et al. (2021a) distinguish two types of approaches to annotation: descriptive and prescriptive. Descriptive annotations encourage subjectivity of annotators (where inconsistency is not an issue), while prescription instructs annotators to strictly follow carefully defined criteria (the less subjectivity, the better). Accordingly, they require definitions that are either more open to interpretation or that are more specific as to what falls under the phenomenon under investigation (in our case *hate speech*).

Most datasets we explored are built to serve as training data for discrete classification identifying if a message contains hate speech (or another form of abusive language). This requires consistently annotated data.[3] As such, most existing datasets should ideally follow a prescriptive paradigm. However, many use definitions that can introduce unintended forms of subjectivity leading to problematic forms of inconsistency.

[1] We emphasize that the focus of this definition is on textual *hate speech*. Introducing other modalities (e.g. images or sound) adds other layers of complexity.

[2] We limit our overview to English for reasons of space, but plan to apply the criteria to all sets mentioned on the site in the future.

[3] We found a notable exception in Ousidhoum et al. (2019) following a *descriptive* approach, who aim to assess how people view and react to *hate speech*. Röttger et al. (2021a) give an overview of abusive language datasets and how they correspond to the annotation paradigm of prescriptive vs. descriptive based on the definition.

As mentioned in Section 1, even in the most prescriptive scenario of all, criminal law, experts may differ in their judgment. A certain level of disagreement is thus inevitable due to the nature of the task. This should nevertheless be limited to borderline cases and the definition should make explicit where this borderline is situated, e.g. should it include all potentially harmful messages to create a pleasant online environment for users or focus on the most extreme cases that potentially break the law?

While borderline cases are inevitable, there are also clear cases where there is wide agreement on a message being an example of hate speech, or contrarily benign without any signal that could possibly be problematic. In general, the fact that there is disagreement on a specific example can be valuable information (Aroyo and Welty, 2015). A system trained on data that captures disagreement could for instance reflect the perception of various annotators (e.g. providing scores that reflect how many annotators would consider an utterance to constitute hate speech). Subjectivity is a strength in this scenario, but the racial bias reported by Sap et al. (2019) would still be problematic. It therefore remains desirable to raise the annotators' awareness of their biases. This would not be the case if the goal of annotating would be to investigate annotator bias rather than creating data for training or evaluating a system. Here, influence on annotators should be kept to a minimum. These examples illustrate that it is important to make conscious decisions as to where subjectivity is desired in annotations and to clearly specify which criteria annotators should not deviate from.

2.3 Defining Hate Speech

There is large variation in current NLP definitions and datasets. This begins with the inconsistent usage of terms. *Abusive* and *offensive language* are examples of terms that have been used to express the same or similar concepts (Schmidt and Wiegand, 2017; Talat et al., 2017; Fortuna et al., 2020; Madukwe et al., 2020). Talat et al. (2017) introduce a typology that aims to further specify types of abusive language by distinguishing between (1) explicit and implicit and (2) directed or generalized forms. We limit ourselves to *hate speech*. We pose that the relation of the negativity of an abusive utterance to a target's (membership of a) specific group is a defining characteristic of *hate speech*.

Even while working on the same phenomenon, there are several (subtle) differences in work addressing *hate speech*. Some datasets have a broader understanding of *hate speech*, e.g. Davidson et al. (2017) take a(n unintended) descriptive approach by not defining potential targets. Fortuna et al. (2020) contrast the more vague definition of Davidson et al. (2017) to the explicit list in Talat et al. (2017), who intentionally focus on a more narrow phenomenon covering only racism and sexism. Fortuna et al. (2020) confirm that different datasets “provide their own flavor of hate speech" (Fortuna et al., 2020, p. 6782).

Varying and vague definitions can lead to inconsistencies that can (unknowingly) be problematic (Madukwe et al., 2020). For instance, users may have a different expectation from a dataset than what its annotations actually cover. Ensuring that datasets are used and created appropriately starts with awareness. Therefore, we introduce *hate speech criteria* (detailed in Section 3) that can be used to construct (prescriptive or descriptive) definitions with annotational guidelines. Individual steps can be adapted depending on the task. Definitions can support a broader or more narrow focus. They can try to leave subjectivity to a minimum or explicitly keep specific aspects underspecified to collect multiple perspectives. Clear definitions can address some challenges around *hate speech* identification, but not all. We elaborate on remaining open issues, such as influence of individual annotators in Section 5.

Our proposal resembles prior work by Kennedy et al. (2022). They translate their definition into a hierarchical coding typology that is used to annotate their *hate speech* dataset. They also use insights from legal (Germany, Australia, The Netherlands, and other countries), sociology, and psychology disciplines. Like us, they point out that *hate speech* is treated differently per country and recognize the importance of having a negative reference relating to (membership of) a group in the utterance. Fortuna and Nunes (2018) discuss differences of hate speech definitions between different sources[4] and recognize different dimensions that are mostly present: having a target, inciting hate or violence, to attack or diminish, and humor having a special status. Zufall et al. (2020) introduce a schema to assess if utterances are hate speech according to the EU law. In contrast to these works, we take a broader approach and present criteria to construct

[4] Platforms, code of conducts, and one scientific paper.

definitions that fit a diverse set of operationalizations according to the desired research. Our criteria can thus support the same definitions Zufall et al. (2020) cover, but is also wide enough to support other types of definitions. In addition, we introduce new aspects that are essential to *hate speech*, such as the dominance of a group and perpetrator characteristics to our criteria.

We furthermore go beyond these prior studies in that we provide an overview of existing English NLP datasets that address *hate speech*. This overview is powered by the dimensions provided in our criteria, which are complementary to the aspects introduced in the typology by Talat et al. (2017), who cover abusive language in general. Their typology does not focus on definitional *hate speech* dimensions but captures the (dis)similarities between different types of abusive language. As mentioned above, they distinguish between abuse directed at an individual or generally addressing a target group and between implicit or explicit abuse. Our approach relates to that typology as follows. Their examples of generalized abusive language would typically fall under *hate speech*. Speech directed at individuals also falls under *hate speech* if there is direct evidence for the abuse being related to group membership. We add specifications on potential targets, group dominance, perpetrator information and the effect of the message which can encapsulate both implicit and explicit *hate speech*.

3 Proposed Hate Speech Criteria

This section provides our proposed procedure to define *hate speech*. As outlined above, we follow the view that hate speech is characterized by problematic statements that are related to a target's (presumed) membership of a specific group. Starting from this assumption, we propose the following criteria, represented in Figure 1, to define the scope of *hate speech*:

1. Identify the target group(s).

2. Specify the social status of the target group

3. Consider properties of the perpetrator

4. Identify the type of negative reference (in relation to the target) present.

5. Identify the potential effects/consequences of the utterance.

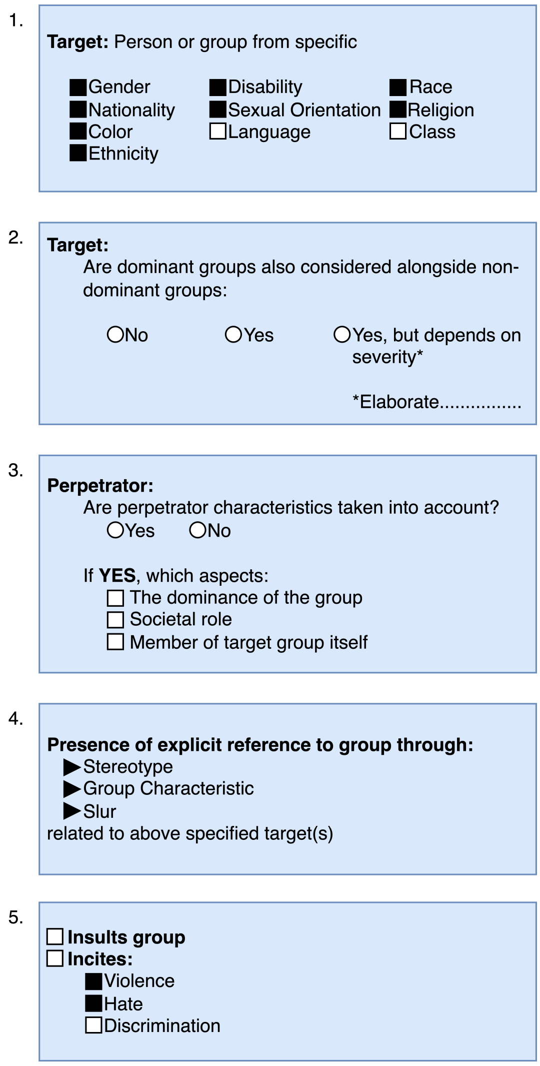

Figure 1: Our proposed *Hate Speech Criteria* to support modular task-specific definition and annotation guidelines construction.

Per step, we indicate the considerations that should be taken into account. These can differ depending on the task, but certain elements are standard across many definitions found in NLP and are also generally supported by law. We call these cases *standard* cases and indicate them in our figure with a filled checkbox. Other facets which are known to be considered in existing definitions (but not all) are *optional* cases and these are left unfilled. This corresponds to how European countries have defined *hate speech*, with some target groups being more common and/or obliged by EU law and other target groups differing among member states (Commission et al., 2021). The options can be adjusted in any way the use case requires: one can extend the definition, narrow it down to investigate a specific form of hate speech or purposely leave a component underspecified. It is specifically possible to work with multiple definitions that apply in different legal or social contexts (e.g. being more lenient to what is allowed in artistic context or being more protective towards users on a social platform). Note that the criteria do not intend to distinguish different forms of *hate speech*, but allow researchers to define or distinguish a specific form themselves when necessary for a task. An example of applying the criteria to a task is given in Appendix B.

3.1 Considered Targets

The inclusion of specific target groups depends heavily on the task (e.g. women- and immigrants-focused (Basile et al., 2019) or racism- and sexism-focused (Talat and Hovy, 2016)). For instance, a law-supporting detection system in Belgium would also consider *language*[5] a basis of a group, while that would not be the case in the Netherlands.[6] Thus, the first step in defining the scope under consideration is to specify which target groups are being considered for your task. In Figure 1, the most common target groups are indicated as the **standard** groups. The list of possible characteristics is not exhaustive and others that historically have been disparaged can be added. Vice-versa, a study may focus on a subset of these groups. Which specific target groups people consider potential victims of hate speech can furthermore be the topic of descriptive research. In this case, it should be defined that this is intentionally left unspecified.

3.2 Target Group Dominance

An important distinction that can be made in the target group is the dominance of a group in society, depending on where the model will be deployed. We define a dominant (cultural) group as a group whose members are (possibly without them being aware) positively privileged (Razzante and Orbe, 2018), unstigmatized (Rosenblum and Toni-Michelle, 2000), and generally favored by societal institutions (Marger, 1997). Hateful sentences against non-dominant groups can be far more consequential than those addressing groups that are in power and can control the narrative. As such, objectionable speech against (an individual from) a dominant group does not necessarily have to be considered *hate speech*. While for some tasks this distinction would be sensible to make, the law does not always make it, e.g. in The Netherlands (Van Noorloos, 2014b).

We ask the question if the task at hand also takes the dominant group into account as a potential target of *hate speech*. There are three different options: The option *no* excludes all forms of negative speech addressed at the dominant group from consideration. The option *yes* does not distinguish between targets from the dominant group and other targets. The third option assumes that utterances targeting the dominant group can be *hate speech*, but under stricter conditions. For instance, a definition may exclude the possibility of discriminating against the dominant group, but would consider calls for violence against them *hate speech*.

3.3 Speaker/Perpetrator

The third distinction we propose is considering perpetrator characteristics (Geldenhuys and Kelly-Louw, 2020). It should be made explicit whether, for a particular task, it matters who the perpetrator is. Because it is a common scenario in NLP that only text is available and the background of a speaker cannot be determined, there are **no standard** aspects to consider here. We describe how speaker characteristics may be taken into account for those scenarios where they can be retrieved. For instance, a person uttering possible *hate speech* against their own group may be "exempted". It is also important to consider what such a speaker is doing with their utterance. If they are "re-appropriating"[7] speech to reject the negative statement (Galinsky et al., 2013), that would not be considered *hate speech* while if the intention is justification, it would be. Additionally, the societal role of the perpetrator may play a role: a person in a powerful position saying something derogating can be much more harmful than an average person saying the same thing, e.g. a CEO of a tech company making derogatory remarks about female engineers. In contrast, e.g. artists can be given more freedom due to artistic expression. Some

[5] Belgian Criminal Code: Articles 377bis, 405quater, 422quater, 438bis, 442ter, 453bis, 514bis, 525bis, 532bis, and 534quater https://www.ejustice.just.fgov.be/cgi_loi/change_lg.pl?language=nl&la=N&cn=1867060801&table_name=wet

[6] Dutch Criminal Code: Articles 137d and 137e https://wetten.overheid.nl/BWBR0001854/2022-03-01

[7] *Reappropriate:* "to take possession for oneself that which was once possessed by another, and we use it to refer to the phenomenon whereby a stigmatized group revalues an externally imposed negative label by self-consciously referring to itself in terms of that label." - Galinsky et al. (2003)

countries, e.g. The Netherlands also allow more space for politicians:[8] Statements that contribute to the political debate are given more protection in lieu of freedom of expression, but remarks may not infringe other rights.[9]

3.4 Types of References to Target Groups

Hate speech is a specific kind of abuse that is characterized by a negative reference that is either aimed at a target group or *explicitly* related to membership of a target group. We thus differentiate between negative behavior toward someone from a potential target group from negative behavior *because of* someone's membership of a target group. For illustration, *"They should lock you up!"* clearly is a problematic message due to its threatening nature. It would nevertheless not be considered hate speech as there is no explicit reference to the individual being a part of a targeted group, even if they are in fact a member of such a group. Now, if we change it to *"They should lock you up, SLUR!"* where the slur specifically targets a group, this *would* be considered hate speech, as the slur clearly signals a relation between the threat and the group the target belongs to.

We explicitly state that the text should contain one (or more) of the following: (i) a stereotype (ii) a group characteristic (this can be the group itself as well) or (iii) a slur that is connected to the target groups specified in the first block. This is the only step where all the references provided are **standard** and cannot be optional. Only if the addressed task enables using more contextual clues while annotating then the reference may be found in a larger context (e.g. another tweet in the thread), but some evidence for the direct link between group membership and the abuse must be present. If there is no larger context, then reliance should be only on the present text.

3.5 Potential Consequence of Utterance

The last step in setting up the scope of *hate speech* is evaluating its strength and potential effects. The actual effects need not be proved, also not in criminal law. However, the words need to be liable to incite to hatred, violence, discrimination or to insult (Van Noorloos, 2014a). Most definitions consider inciting violence and hate as *hate speech*, as these consequences make *hate speech* stand out from other offensive expressions. These two incitements are **standard** cases. Additional broader potential consequences can also be considered, such as inciting discrimination, or a general insult toward a group. The latter is specifically recognized by Dutch law. Which possible consequences should be taken into account depends on the severity of *hate speech* the task should address. It furthermore depends on the context wherein the narrative exists. Is there a relation to a threatening historical situation? Does the uttering call for exclusion of particular target groups? Furthermore, a threat can be implicitly present, i.e. "What should we do with your *stereotypical object*?". While there is not an explicit *call* for it, violence *is* implied: destruction. The threat lies in its potential consequences. It is important to understand the implications and where the possible violence or hate stems from in a statement. Once this is understood, one may decide if different consequences, depending on severity, should apply to different targets or not.

4 Overview of Definitions and Datasets

To highlight the differences between existing datasets and to comprehend what kind of tasks they would fit, we present an overview of widely-used datasets based on their definitions. Our scope is restricted to all English datasets found on `hatespeechdata.com` that tackle *hate speech*, since it has the most datasets and a variety of definitions.[10] The overview can be found in Table 1, where we indeed observe this variation.

For each step in our criteria, we indicate if it is explicitly specified in the definition or not. There is a difference between defining *hate speech* and specifying a particular focus. We follow the descriptions of the annotations in the dataset for our classification. In cases where *only* a definition is given, we assume that that is the focus of the dataset as well, unless stated otherwise (e.g. in Talat and Hovy (2016); Basile et al. (2019)). An ✗ signals that the aspect is unspecified in the definition, therefore it is still possible that the specifics are present in the dataset but this cannot be guaranteed. E.g.

[8] `https://mensenrechten.nl/nl/vrijheid-van-meningsuiting`

[9] An example of a politician violating this freedom: `https://uitspraken.rechtspraak.nl/inziendocument?id=ECLI:NL:RBAMS:2021:7392`, `https://www.politico.eu/article/dutch-mep-guilty-anti-semitism-holocaust/`

[10] We leave out Sarkar and KhudaBukhsh (2020) as it presents a challenge for hate speech detection models but does not address the task itself. We include Zufall et al. (2020) to illustrate that our criteria also fits a legal perspective of *hate speech*.

	T	ND	P	Explicit Ref	Effects/Consequences
Talat and Hovy (2016)	✓	✓	✗	Stereotype & Slur	Insult, Violence, Hate, *Other*
ElSherief et al. (2021)	✓	✗	✗	Slur & Group Characteristics	Insult, Violence, Hate, Discrimination, *Other*
Kennedy et al. (2022)	✓	✗	✗	Slur, Group Characteristics, & Stereotypes	Violence or Hate
Basile et al. (2019)	✓	✓	✗	✗	*Other*
Kirk et al. (2021)	✓	✗	✗	✗	Discrimination, *Other*
Founta et al. (2018)	✓	✗	✗	✗	Insult, Hate, *Other*
Mandl et al. (2019)	✓	✗	✗	Stereotypes & Group Characteristics	✗
Mollas et al. (2020)	✓	✗	✗	✗	Insult, Violence
Zufall et al. (2020)	✓	✗	✗	✗	Violence, Hate
ElSherief et al. (2018)	✓	✗	✗	✗	*Other*
Gao and Huang (2017)	✓	✗	✗	✗	*Other*
Qian et al. (2019)	✓	✗	✗	✗	*Other*
Ribeiro et al. (2018)	✓	✗	✗	✗	Violence, *Other*
Röttger et al. (2021b)	✓	✗	✗	*Other*	✗
Chung et al. (2019)	✓	✓	✗	✗	✗
Fanton et al. (2021)	✓	✓	✗	✗	✗
Mathew et al. (2021)	✓	-	✗	✗	✗
Davidson et al. (2017)	✗	✗	✗	✗	Insult, Violence, Hate, *Other*
de Gibert et al. (2018)	✗	✗	✗	*Other*	✗
Ousidhoum et al. (2019)	✓	✗	✗	✗	✗

Table 1: Overview of existing datasets according to the hate speech criteria that we propose. **T:** target groups specified, **ND:** considering only non-dominant groups specified, **Explicit Ref**: explicit reference specified, if yes; which ones, **Effects/Consequences:** effects/consequences specified, if yes; which ones. Per paper we indicate for each aspect if they are present in the definition or focus.

several datasets might only consider non-dominant groups but do not explicitly state so, or when left unspecified it is unclear which explicit references to the group are always present.

Under column **T** we see if there are specific target groups in the definition. There are very few datasets that do not explicitly mention their target groups (Davidson et al., 2017; de Gibert et al., 2018). Although there is some overlap in groups between different datasets, there are also (subtle) differences. Due to this variety, we provide an overview of different target groups covered per dataset in Appendix A.

For the second step, most definitions do not mention anything about dominance. Mathew et al. (2021) are the only ones to mention *Caucasian* as a target group, which we mark with a '-' to signal that this paper is explicit about not restricting itself to non-dominant groups. Most papers with a ✓for **ND** specifically define their targets to apply to non-dominant groups only (Chung et al., 2019; Basile et al., 2019; Fanton et al., 2021), with the exception of Talat and Hovy (2016), who mention *minorities*.

None of the datasets mention taking perpetrator characteristics into account (column **P**). Similarly, the explicit references are left unspecified in most datasets' definitions (column **Explicit Ref**). This means that for such datasets it cannot be guaranteed whether explicit references are present, nor whether they include specifications as to which ones. Looking at **Effects/Consequences**, *violence* and *hate* occur the most. Other terminology for negative relations and effects/consequences widely differs and the interpretation with respect to our criteria can be subjective (e.g. is 'humiliate' a form of discrimination, an insult or something else?), we mark such terminology as *Other*.

The idea behind the overview is that it can illustrate the need for future datasets that specify their aspects more explicitly and aids in deciding which dataset is suitable for a specific task. For instance, if a task requires a dataset that guarantees a focus on non-dominant groups, then column **ND** can easily point to the datasets that fit this prerequisite explicitly. If the dominance being specified is not very important but the presence of an explicit reference of a negative relation like a slur is, then Mandl et al. (2019) could be fitting for the task. If, in addition, incitement of hate and violence is essential, then Kennedy et al. (2022) should be considered. If dominance is important as well, one might consider to further annotate samples from these sets that are labeled as *hate speech*, saving the time to separate 'clean' messages. In combination with the overview of which datasets cover which target groups, we believe these outlines to be helpful for identifying useful datasets.

5 Discussion

The presented *hate speech* criteria aim to include those aspects of *hate speech* needed to arrive at clearer definitions and to provide better annotation guidelines, while supporting a wide range of use cases. We are aware, however, that they do not provide a magic solution to all challenges around this complex phenomenon. In this section we briefly discuss (1) possible extensions, (2) possible further specifications and (3) challenges that a clear definition cannot (fully) address on its own.

We tried to create an extensive overview of relevant aspects, but are well aware that we may have missed things. Moreover, *hate speech* is strongly connected to culture and what is perceived as *hate speech* may change. The criteria can thus be **extended** to cover new target groups, more perpetrator characteristics, or additional potential consequences. This particularly holds for the fourth step: *Types of References to Target Groups*. We maintain that *evidence* of the abuse being related to (assumed) group membership is a requirement, but researchers may decide that other clues can also serve as possible evidence for an utterance being *hate speech*. These clues may include the history of a certain perpetrator, who in the past has uttered instances of *hate speech* multiple times. As mentioned, the evidence may also come from e.g. what the utterance is responding to.

The typology of Talat et al. (2017) is not included in our criteria. We explained in Section 2 how our criteria relate to this typology. It is however straightforward to add **further specifications** to a *hate speech* definition created through our criteria. Note that, even though our criteria relies on the clear presence of group characteristics, stereotypes, and/or slurs, they do not exclude implicit forms of abuse, especially since we also evaluate the potential consequences e.g.: *"Everything was quite ominous with the train accident. Would like to know whether the train drivers were called StereotypicalName1, StereotypicalName2 or StereotypicalName3 #RefugeeCrisis"* (Benikova et al., 2017). Here, the stereotypical names indicate that this falls under *hate speech*. Utterances like "*white revolution is the only solution*"(ElSherief et al., 2021) may seem problematic due to the lack of an explicit slur, stereotype or target group characteristic. It nevertheless provides direct evidence, since "white" implies that the revolution would be against non-white and the violent nature of the threat.

A clear definition can help avoid inconsistencies and unwanted forms of subjectivity in the data, but it cannot address all challenges involved in determining whether an instance exhibits *hate speech*. First, we mentioned multiple times that some form of **subjectivity** remains inevitable when dealing with *hate speech*. Though people will always differ as to where they draw the line, instructions on the level of severity that should be included with illustrating discussions can be helpful. Even for the seemingly clear case of inciting violence, which is a core aspect, there is a vast difference between uttering *"Throw tomatoes at them!"* and an actual life-threatening *"Gun them down!"*. The class of group insult in particular can include a large variety, from merely unkind statements *"Women really have a horrible sense of fashion with their white sneakers!"* to insults that question people's capabilities or attack someone's morals. Questioning a groups capabilities can lead to discrimination, especially when uttered by people with authority or in power. The remark *"I'm not sexist but female comedians just are not funny!"* (Shvets et al., 2021) may seem relatively harmless when coming from a tweeter with few followers, but when coming from an influential critic or the president of a comedians' union, it can actively harm women's careers. Attacks on a group's morals can also have an impact beyond merely insulting. For instance, saying that a non-dominant group are leeches can incite hate or lead to violent ramifications.

Explanations that illustrate the potential affect on different targets can help annotators to determine the severity of a specific statement and may help them to make more systematic decisions on where to draw the line. This leads to the second challenge that a definition by itself cannot solve: **annotator bias**. Explanations and training may help annotators to tackle their bias and may make them more sensitive to more subtle attacks to groups they are not part of, but the affect of an annotator's background will not be completely eliminated. Our criteria are meant to support creating a definition and guidelines. They do not mention gathering annotator information, because we believe that the definition crafted for a specific task does not change based on annotator information. However, we want to emphasize that it is essential that annotator demographics are taken into account and that the goal of the task should be kept in mind when establishing annotators' background. E.g. if the target group considered for the task is *Gender*, it is of utmost importance to have annotators that can capture experiences of all genders. In general, it is vital to include members of potential target groups, since they are more likely to pick up on subtleties.

A third issue that can only partially be addressed by means of a clear definition lies in the relation between *hate speech* and **freedom of speech**. As mentioned, *hate speech* can create unsafe environments that hamper freedom of speech. At the same time, opinions can differ regarding whether specific remarks are harmful or should be allowed because they are part of an important debate (and where marking them as *hate speech* would hamper freedom of speech). Law has to clearly define what is punishable *hate speech* and what on the other hand should be protected by freedom of speech. In Dutch law, a distinction is made for e.g. public debates where politicians are given more (but not unlimited) space for *controversial* statements. Another example is that Dutch law protects members of a religion from problematic utterances (which is prohibited in all EU countries), but leaves ample space for negative statements about specific religions (decriminalized in many Western countries) (Van Noorloos, 2014a). In the context of creating a safe environment for discussion, this distinction between attacking a religion or people can be hard to make sometimes and there are cases where it does not seem to make sense. The example *#BanIslam* from Talat and Hovy (2016), for instance, might be aimed at religion and not at people, but it can clearly be harmful to Muslims and it is hard to see how such a hashtag would contribute to a useful public debate or discussion. In this context, both the potential harm and added value of statements to a debate should be taken into account. Though the example of *#BanIslam* seems clear, it is easy to imagine that it is not always straightforward to make this call.

The challenges mentioned above show that a clear definition may be a good starting point, but cannot solve everything. We provided examples that illustrate the complexity, but full discussions would merit individual papers on each of these topics. A final limitation we want to point out is that our overviews are currently limited to English datasets. Our framework leaves variables related to linguistic properties or cultural aspects open and can thus be easily applied to datasets covering other languages.

6 Conclusion

We presented modular criteria to construct definitions and annotator guidelines to address *hate speech*. These steps include aspects that have, to our knowledge, not been prominent before. We propose five components for defining hate speech: (1) identifying the target group(s), (2) specifying the consideration of dominant groups, (3) considering perpetrator characteristics, (4) finding explicit negative reference(s) of the target(s), and (5) identifying the potential consequences/effects. Based on the task at hand, the definition can be modified as, depending on the application for which *hate speech* is addressed, a different description may be needed. This ties into how strictly each specific aspect needs to be defined as well: do we need annotations that are as consistent as possible (prescriptive) or do we want to investigate diversity in perspectives on this particular aspect (descriptive)?

We provided an overview of a large variety of English *hate speech* datasets based on the dimensions that are present in our criteria. We hope that the criteria and discussions in this paper will motivate NLP researchers working on *hate speech* to critically think about the tasks they are addressing and evaluate how fitting current definitions and datasets are for their task. The overview can then help select the most suitable datasets that can either be directly used or used as starting points that serve the task after adding further specifications. We particularly hope that the discussion in this work will help those working on new datasets to take these aspects into account from the start.

Acknowledgements

This research was (partially) funded by the Hybrid Intelligence Center, a 10-year programme funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research. We would additionally like to thank the reviewers for providing us with valuable feedback that has helped improving this paper.

References

Hala Al Kuwatly, Maximilian Wich, and Georg Groh. 2020. Identifying and measuring annotator bias based on annotators' demographic characteristics. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 184–190, Online. Association for Computational Linguistics.

Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24.

Md Rabiul Awal, Rui Cao, Roy Ka-Wei Lee, and Sandra Mitrović. 2020. On analyzing annotation consistency in online abusive behavior datasets. *arXiv preprint arXiv:2006.13507*.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Darina Benikova, Michael Wojatzki, and Torsten Zesch. 2017. What does this imply? examining the impact of implicitness on the perception of hate speech. In *International Conference of the German Society for Computational Linguistics and Language Technology*, pages 171–179. Springer.

Carmen Cervone, Martha Augoustinos, and Anne Maass. 2021. The language of derogation and hate: Functions, consequences, and reappropriation. *Journal of language and social psychology*, 40(1):80–101.

Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. 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*, pages 2819–2829, Florence, Italy. Association for Computational Linguistics.

European Commission, Directorate-General for Justice, Consumers, P Ypma, C Drevon, C Fulcher, O Gascon, K Brown, A Marsavelski, and S Giraudon. 2021. *Study to support the preparation of the European Commission's initiative to extend the list of EU crimes in Article 83 of the Treaty on the Functioning of the EU to hate speech and hate crime : final report*. Publications Office.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11.

Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate speech dataset from a white supremacy forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 11–20, Brussels, Belgium. Association for Computational Linguistics.

Mai ElSherief, Shirin Nilizadeh, Dana Nguyen, Giovanni Vigna, and Elizabeth Belding. 2018. Peer to peer hate: Hate speech instigators and their targets. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Margherita Fanton, Helena Bonaldi, Serra Sinem Tekiroğlu, and Marco Guerini. 2021. Human-in-the-loop for data collection: a multi-target counter narrative dataset to fight online hate speech. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3226–3240, Online. Association for Computational Linguistics.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51:1 – 30.

Paula Fortuna, Juan Soler, and Leo Wanner. 2020. Toxic, hateful, offensive or abusive? what are we really classifying? an empirical analysis of hate speech datasets. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6786–6794, Marseille, France. European Language Resources Association.

Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.

Adam D Galinsky, Kurt Hugenberg, Carla Groom, and Galen V Bodenhausen. 2003. The reappropriation of stigmatizing labels: Implications for social identity. In *Identity issues in groups*, volume 5, pages 221–256. Emerald Group Publishing Limited.

Adam D Galinsky, Cynthia S Wang, Jennifer A Whitson, Eric M Anicich, Kurt Hugenberg, and Galen V Bodenhausen. 2013. The reappropriation of stigmatizing labels: The reciprocal relationship between power and self-labeling. *Psychological science*, 24(10):2020–2029.

Lei Gao and Ruihong Huang. 2017. Detecting online hate speech using context aware models. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 260–266, Varna, Bulgaria. INCOMA Ltd.

Judith Geldenhuys and Michelle Kelly-Louw. 2020. Demystifying hate speech under the pepuda. *Geldenhuys J and Kelly-Louw M" Demystifying Hate Speech underthe PEPUDA" PER/PELJ*.

Vebjørn Isaksen and Björn Gambäck. 2020. Using transfer-based language models to detect hateful and offensive language online. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 16–27, Online. Association for Computational Linguistics.

Brendan Kennedy, Mohammad Atari, Aida Mostafazadeh Davani, Leigh Yeh, Ali Omrani, Yehsong Kim, Kris Coombs, Shreya Havaldar, Gwenyth Portillo-Wightman, Elaine Gonzalez, et al. 2022. Introducing the gab hate corpus: defining and applying hate-based rhetoric to social media posts at scale. *Language Resources and Evaluation*, pages 1–30.

Svetlana Kiritchenko, Isar Nejadgholi, and Kathleen C Fraser. 2021. Confronting abusive language online: A survey from the ethical and human rights perspective. *Journal of Artificial Intelligence Research*, 71:431–478.

Hannah Rose Kirk, Bertram Vidgen, Paul Röttger, Tristan Thrush, and Scott A Hale. 2021. Hatemoji: A test suite and adversarially-generated dataset for benchmarking and detecting emoji-based hate. *arXiv preprint arXiv:2108.05921*.

Kosisochukwu Madukwe, Xiaoying Gao, and Bing Xue. 2020. In data we trust: A critical analysis of hate speech detection datasets. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 150–161, Online. Association for Computational Linguistics.

Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. 2019. Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. In *Proceedings of the 11th forum for information retrieval evaluation*, pages 14–17.

Martin Marger. 1997. *Chapter 5: Foundations of the American Ethnic Hierarchy: Anglo-Americans and Native Americans*, 4th edition, page 146–169. Wadsworth, Belmont.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14867–14875.

Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. 2020. Ethos: an online hate speech detection dataset. *arXiv preprint arXiv:2006.08328*.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 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)*, pages 4675–4684, Hong Kong, China. Association for Computational Linguistics.

Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. 2019. A benchmark dataset for learning to intervene in online hate speech. 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)*, pages 4755–4764, Hong Kong, China. Association for Computational Linguistics.

Robert J Razzante and Mark P Orbe. 2018. Two sides of the same coin: Conceptualizing dominant group theory in the context of co-cultural theory. volume 28, pages 354–375. Oxford University Press.

Manoel Ribeiro, Pedro Calais, Yuri Santos, Virgílio Almeida, and Wagner Meira Jr. 2018. Characterizing and detecting hateful users on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.

Karen E Rosenblum and C. Travis Toni-Michelle. 2000. *The Meaning of Difference: American Constructions of Race Sex and Gender Social Class and Sexual Orientation*, 2nd edition. McGraw-Hill, Boston.

Paul Röttger, Bertie Vidgen, Dirk Hovy, and Janet B Pierrehumbert. 2021a. Two contrasting data annotation paradigms for subjective nlp tasks. *arXiv preprint arXiv:2112.07475*.

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021b. HateCheck: Functional tests for hate speech detection models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 41–58, Online. Association for Computational Linguistics.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Rupak Sarkar and Ashiqur R KhudaBukhsh. 2020. Are chess discussions racist? an adversarial hate speech data set. *arXiv preprint arXiv:2011.10280*.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Alexander Shvets, Paula Fortuna, Juan Soler, and Leo Wanner. 2021. Targets and aspects in social media hate speech. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 179–190, Online. Association for Computational Linguistics.

Zeerak Talat. 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*, pages 138–142, Austin, Texas. Association for Computational Linguistics.

Zeerak Talat, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84, Vancouver, BC, Canada. Association for Computational Linguistics.

Zeerak Talat and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Marloes Van Noorloos. 2014a. Criminalising defamation of religion and belief. *European Journal of Crime, Criminal Law and Criminal Justice*, 22(4):351–375.

Marloes Van Noorloos. 2014b. The politicisation of hate speech bans in the twenty-first-century netherlands: Law in a changing context. *Journal of Ethnic and Migration Studies*, 40(2):249–265.

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

Jeremy Waldron. 2012. The harm in hate speech. In *The Harm in Hate Speech*. Harvard University Press.

Frederike Zufall, Marius Hamacher, Katharina Kloppenborg, and Torsten Zesch. 2020. A legal approach to hate speech: Operationalizing the eu's legal framework against the expression of hatred as an nlp task. *arXiv preprint arXiv:2004.03422*.

A Overview of Target Groups in Datasets

Since many of the datasets have varied target groups that are taken into consideration, we present an overview of the target groups that are mentioned in their definitions, or are explicitly stated to be their focus, in Table 2. Due to different terminology used for related concepts, we use umbrella terms for the distinct categories and indicate the precise terms if those categories are present in the dataset (e.g. health concerns, disease, and disability grouped under health). Under **Other** we illustrate target groups that do not fit the other categories and do not occur enough to be specified by themselves. Furthermore, we also indicate if definitions keep the target groups open to unspecified ones by using wordings like "groups such as ..." (e.g. Kirk et al. (2021); Mandl et al. (2019)).

Datasets that do not have any target groups indicated, as can be seen in Table 1, are left out from this overview.

When a specific focus is mentioned (e.g. Chung et al. (2019); Basile et al. (2019)), instead of using the umbrella terms, we use the exact targets as mentioned by the paper. Moreover, when both *Gender* and *Gender Identity* are considered by a dataset, this is indicated as *Gender (Identity)*.

	Gender	Origin	Religion	Sexual Orientation	Health	Other
Talat and Hovy (2016)	Gender	Race				
ElSherief et al. (2021)	Gender (Identity), Sex	Race, Ethnicity, Nationality	Religion	Sexual Orientation	Disability, Disease	Age
Kennedy et al. (2022)	Gender	Race, Ethnicity, Nationality, Regionalism	Religion, Spiritual Identity		Mental, Physical Health	Ideology, Political Identification
Basile et al. (2019)	Women	Immigrants				
Kirk et al. (2021)	Gender	Race, Ethnicity, Nationality, Color, Descent	Religion			"Other identity factor"
Founta et al. (2018)	Gender	Ethnicity, Race	Religion	Sexuality	Disability	"Attributes such as"
Mandl et al. (2019)	Gender	Race		Sexual Orientation	Health Condition	Political Opinion, Social Status, "or similar"
Mollas et al. (2020)	Gender	Race, National Origin	Religion	Sexual Orientation	Disability	
Zufall et al. (2020)		Race, Colour, Descent, National or Ethnic Origin	Religion			
ElSherief et al. (2018)	Gender, Sex	Race, Ethnicity, National Origin	Religion	Sexual Orientation	Disability, Disease	
Gao and Huang (2017)	Gender	Ethnicity		Sexual Orientation		"Facet of identity"
Qian et al. (2019)	Gender (Identity), Sex	Race, Ethnicity, National Origin, Caste	Religion	Sexual Orientation	Disease, Disability	
Ribeiro et al. (2018)	Gender (Identity)	Race, Ethnicity, National Origin	Religion	Sexual Orientation	Disability, Disease	Age
Röttger et al. (2021b)	Women, Trans people	Black people, Immigrants	Muslims	Gay people	Disabled people	
Chung et al. (2019)			Islamophobia			
Fanton et al. (2021)	Women	People of Color, Romani, Migrants	Jews, Muslims	LGBT+	Disabled people	Overweight people
Mathew et al. (2021)	Gender	Race, Indigenous, Refugee, Immigrant	Religion	Sexual Orientation		
Ousidhoum et al. (2019)	Gender	Origin	Religion	Sexual Orientation	Special Needs	

Table 2: Overview of target groups. Each column represents a type of target, under which we indicate the specific targeted group per dataset. An empty cell indicates that the target group type was not mentioned in the definition/focus.

B Example of Applying the Criteria to a Task

To showcase the utility of the proposed criteria to create a definition and associated annotation guidelines, we will apply the criteria to a prescriptive and descriptive scenario. For reasons of simplicity, we assume the questions are applied to texts that contain some form of abuse (where we use *abuse* as an overarching term which includes any form of *hate speech* as well as other forms potentially harmful content).

B.1 Prescriptive scenario

Consider the following scenario where we want to create a definition for a task that takes down *hate speech* on a social media platform that goes against the Dutch law for *hate speech*.[11] We initially want annotations in a prescriptive setting: as the addressed task concerns the law, we strive to reduce subjectivity to a minimum.

For each step, we fill in what is specified according to the Dutch law. Correspondingly, the target groups considered are: race, religion or philosophy of life, gender, hetero- or homo-sexuality, and physical and mental disability. As the law does not make a distinction between dominant and non-dominant groups, the task will not either. The step around perpetrator is more complex: the law does not define a distinction for the kind of perpetrator, but does allow for more lenience in a political or artistic context. We require the presence of all the explicit references mentioned in the criteria as all of them are standard cases: stereotype, group characteristic, and slur. Furthermore, the incitement of violence, hate or discrimination is considered. In addition, group insult is also seen as a consequence for all target groups except for gender.

This brings us to the following definition for the task: *For this task, hate speech is defined as language targeted at a person or group based on their race, religion or philosophy of life, gender, hetero- or homo-sexuality, and physical and mental disability and incites violence, hate or discrimination or insults a group on the basis of aforementioned targets, barring gender.*

Then, we transfer this definition using the criteria to annotation guidelines. For each step, we ask the question if the specification is present or not. If any step is not considered, we keep the lack of consideration as a note to prevent personal judgments in such cases as much as possible. If the answer to a question is *yes*, the annotator can proceed to the next question. If an answer is no, the instance is not covered by the task definition of *hate speech* and the annotator can directly label the instance as *not hate speech*.

For this specific task, we ask the following questions to be answered for texts that contain a form of abuse:

1. Does the target (group) belong to one of the following groups: *race, religion or philosophy of life, gender, hetero- or homo-sexuality, and physical and mental disability*?

2. **NOTE:** The target (group) can belong to both non-dominant and dominant groups.

3. Does the text contain an explicit reference to the group (related to above specified target(s)) through a stereotype, group characteristic or slur?

4. Does the text incite violence, hate, or discrimination or group insult? If the text incites violence, hate or discrimination, it should be labeled as *hate speech*. If the text only contains a group insult proceed to the next question.

5. Is the group insult directed at a group based on the following characteristics: *race, religion or philosophy of life, hetero- or homo-sexuality, and physical and mental disability*? If yes, it should be labeled as *hate speech*.

6. If the text contains *hate speech*: Is the speaker an artist or politician making the utterance in a context of their work? Please indicate "political or artistic context" (it can still be hate speech).

Please note that the instructions above literally follow Dutch law and we are aware that these targets does not cover all groups (notably this is a rather limited view on LGBTQ+) and that the exclusion of gender from group insult is debatable. It should also be noted that aspects such as group dominance or who the speaker is can have an influence on a verdict, but applying those subtleties would, in this scenario, be up to judges making a final verdict rather than annotators marking potential violations of the law. For similar reasons, we

[11] Dutch Criminal Code: Articles 137d and 137e https://wetten.overheid.nl/BWBR0001854/2022-03-01.

choose to mark the artistic or political context as a relevant aspect rather than specifying what this would mean for the ultimate decision.

B.2 Descriptive Scenario

Suppose that we want to study what groups various annotators consider potential victims of *hate speech*, including whether they distinguish between dominant or non-dominant groups. We focus on forms of hate speech that incite violence hate or discrimination, leaving the more vague category of group insults out. The questions they should then answer about texts containing some form of abuse are:

1. Is the abuse aimed at a specific target group or a member of such a group?
2. Does the text contain an explicit reference to the group (related to above specified target(s)) through a stereotype, group characteristic or slur?
3. Does the text incite violence, hate, or discrimination?

By making it clear for which aspects the subjectivity from annotators is desired, it is easier to ensure that annotators will deviate from the given instructions for that facet and that other researchers are aware of the variation. We know, for instance, that they did check whether there is an explicit reference to the group. The awareness can help in making an informed decision if the dataset is useful for their task or not. Naturally, the outcome will remain somewhat cluttered by subjective interpretations whether a specific remark could incite, e.g., discrimination. We can however distinguish between these motivations by making annotators answer the question rather than making them label the data. As such, we learn whether their reason for 'no' was related to the specific group being a potential target or to the severity or nature of the abuse.

For reasons of simplicity, we left the criterion of perpetrator information out of this example. Investigating this aspect would probably require a different setup. For instance, first defining *hate speech* as something that incites violence, hate or discrimination, and then asking the following questions:

1. Is the abuse aimed at a specific target group or a member of such a group?
2. Does the text contain an explicit reference to the group (related to above specified target(s)) through a stereotype, group characteristic or slur?
3. Is the speaker a member of the target group?
4. Do you consider this utterance to be *hate speech* based on the definition provided above?

Note that the examples in this appendix are included for purposes of illustration as how definitions may help specifying annotation tasks. We are well aware that providing a good annotation setup, especially for descriptive scenarios, is complex. As many aspects mentioned in this paper, the next step of actually setting up such tasks can merit a paper on its own.