A Typology and Coding Manual for the Study of Hate-based Rhetoric

Brendan Kennedy*, Drew Kogon*, Kris Coombs,
Joe Hoover, Christina Park, Gwenyth Portillo-Wightman,
Aida Mostafazadeh, Mohammad Atari, Morteza Dehghani

University of Southern California

July 18, 2018

1 Introduction

The growth of the Internet as a medium for human communication has increased the visibility of aggressive, attacking, dehumanizing, and potentially dangerous language. The study of hate speech — understanding its causes, its effects on violence and other forms of hate crime, and legal and societal measures for its prevention — has been invigorated and redirected by this increased exposure. In the legal community, this is seen in the debate about censorship and freedom of expression adapting to the challenges posed by online hate speech (e.g. Weinstein, 2018; Gagliardone et al., 2015). In the Natural Language Processing (NLP) community, the prevalence of such undesirable language in online environments has led to a surge in data-driven efforts to detect and predict hate speech (e.g. Warner and Hirschberg, 2012; Olteanu et al., 2018; Davidson et al., 2017; Waseem and Hovy, 2016).

Throughout these developments, the definition of hate speech has been evolving to account for differences in the linguistic content and the intended effects of hate speech. This fluidity is not new: as Sellars (2016) explains, the concept of hate speech has never been fixed and agreed upon. Differences in countries' legal definitions of hate speech have led to divergence in the definitions used by the new wave of hate speech studies. Thus, because a priori, unified conceptions of hate speech do not exist, recent content analyses have partitioned and expanded the areas of research to include abusive language (Nobata et al., 2016), incivility (Anderson et al., 2014), hateful stereotypes (Warner and Hirschberg, 2012), offensive language (Davidson et al., 2017), and personal attacks (Wulczyn et al., 2017). Legal scholars have added to this partitioning by considering the political and societal effects of hate speech, defining "fear" speech (Buyse, 2014), and "dangerous" speech (Benesch, 2012).

Some work has tried to make sense of this outgrowth of hate speech research, providing greater consensus in its definition and operationalization: Schmidt and Wiegand (2017) review methods and data sources from studies which attempt the detection of hate speech, including common findings as the predictiveness of certain types of features; Waseem et al. (2017) construct a typology of abusive language, attempts a more systematic definition of the sub-tasks associated with predicting hate speech; and Warner and Hirschberg (2012) took traditional definitions of hate speech and applied them to real data, observing that much hate speech rhetoric relies on well-known stereotypes.

Through an emphasis on the intention of the speaker, on the role of historical, cultural, ideological, and political context in the rhetorical arguments of hate speech, and on the distinction between language which incites violence versus that which incites hatred, we propose a new definition, typology, and coding procedure for what we call *hate-based rhetoric*. We attempt to justify each by theoretical means and/or accepted standards in the hate speech literature.

Though we attempt to organize and redirect studies of *hate-based rhetoric*, we realize that coding it is non-trivial, to say the least. The contextual factors we mentioned earlier, in addition to the complexity of the rhetorical arguments used in such hate speech, make understanding the intentions of the speaker highly subjective. These issues are magnified by the lack of legal consensus (between countries) as to what constitutes hate speech, and the fact that the motives of those who study it, notably social media companies like Facebook and Twitter, are highly diverse. Lastly, the controversial relationship between hate speech and freedom of speech will always generate detractors of any offered definitions, as they will indubitably leave out instances or include instances incorrectly. Together, these factors are behind the low annotator agreement in coding hate speech (demonstrated by Ross et al., 2017), and make the soundness of proposed definitions and coding procedures all the more important.

Contributions

In our own research, we are motivated to consider the psychological antecedents of hatred, specifically, the role of morality in its language, community dynamics, and relation to aggressive action or violence. The definitions and typologies used in computational studies of hate speech fall short of this mark, however. While they are tuned towards linguistic structure, they are not sufficiently integrated with *context*, or external information, such as historical, political, or cultural references. More generally, existing definitions of hate speech in the literature are not holistic in their view of intent. Our first contribution, in Section 3.1, is the introduction and justification of the concept of *hate-based rhetoric*, a category of language which overlaps with notions of hate speech but is defined by its intent: to dehumanize, oppress, or directly subjugate its target. For example, *hate-based rhetoric* includes such acts as the endorsement of hateful groups or ideology and excludes the use of abusive language which does not include reference to group identity.

Most research which conducts annotation of hate speech — in all of its cat-

egories — lacks theoretical coherence in the development of typology; many projects simply have binary indications of whether or not a document is hateful/abusive/offensive/etc. This annotation style does not help the community to investigate the causes and communicative dynamics of hate speech. Realizing this limitation, recent innovations to the coding of hate speech include the definition and categorization of target populations (Mondal et al., 2017), the introduction of types of "framing" and "speech acts" (Olteanu et al., 2018), and the division between "implicit or explicit" speech (Waseem et al., 2017). In the same vein, our second contribution is to specify a typology of "hate-based rhetoric" (Section 3.2), wherein we borrow from the aforementioned typologies and propose a new, theoretically-oriented distinction between the *incitement to hate* and the *incitement to violence*.

Our categorization of hate speech is similar to that of previous Natural Language Processing (NLP) work such as Warner and Hirschberg (2012) and Waseem and Hovy (2016), which consider the myriad rhetorical devices used to convey hateful attitudes toward an out-group and the role of intention in labeling hate speech. A limitation of such research, however, is that they do not provide a coding manual for the annotation of such rhetoric, nor do they provide a framework amenable to the labeling of language which attacks *any* type of group through any type of linguistic act. Instead, the labeling methods are developed ad hoc and only for particular anti-Semitic (Warner and Hirschberg, 2012) and racist/sexist (Waseem and Hovy, 2016) language. Therefore, in addition to providing a flexible and systematic typology of *hate-based rhetoric*, we provide the first draft of a coding manual (Section 4), which has already been used to train annotators to label social media posts.

2 Background

In our review, we have focused on definitions of hate speech drawn from two general categories: (1) legal definitions, having to do with censorship and the protections of free speech afforded by the state; and (2) definitions used in practice on data-driven analyses of hate speech. To contextualize our view of hate speech and to motivate the need for a new category, *hate-based rhetoric*, we review definitions and perspectives from each of these categories.

Legal Definitions and Perspectives on Hate Speech

As we have alluded to before, legal differences in countries' definitions of hate speech have made the task of applying these definitions to data difficult. Specifically, the United States laws on hate speech are categorically different from other countries, protecting many acts of hate speech which are prohibited in other countries. This narrow definition is adequately summed up by Supreme Court Justice Frank Murphy:

> There are certain well-defined and limited classes of speech, the prevention and punishment of which have never been thought to raise

> a Constitutional problem. These include the lewd and obscene, the profane, the libelous and the insulting or 'fighting' words — those which by their very utterances inflict injury or tend to incite an immediate breach of the peace.[1]

This definition has been held up in court numerous times, including speech that is *prima facie* hateful towards a particular group; these typically enter into a debate about whether the speech uttered constitutes "fighting words" or merely "offensive language". For example, vulgar or offensive words — ones that are not lewd and obscene — are protected under the first amendment and are not seen as fighting words, even if those words include a threat. [2,3,4]

So what are fighting words? The Supreme Court of the United States (SCOTUS) has decided that laws banning acts such as burning a cross on public or private property intended to "arouse anger, alarm or resentment in others" based on "race, color, creed, religion or gender" do not constitute fighting words, but rather the "communication of a political idea". In Justice Scalia's opinion for the Court, he states that:

> Those who wish to use "fighting words" in connection with other ideas — to express hostility, for example, on the basis of political affiliation, union membership, or homosexuality – are not covered [under the First Amendment] [...] the reason why fighting words are categorically excluded from the protection of the First Amendment is not that their content communicates any particular idea, but that their content embodies a particularly intolerable (and socially unnecessary) *mode* of expressing whatever *idea* the speaker wishes to convey [rather than the idea itself]. [5]

Although Scalia clarifies that the court found the actions reprehensible, this interpretation of First Amendment protections means that only speech that would be offensive to any person and that are unrelated to the expression of a political idea can be considered fighting words. In contrast, a statute banning the burning of a cross *with an intent to intimidate* (rather than to express a political idea and not limited to any class of people) was constitutional because the act was a "true threat". This is a term which has never been explicitly defined, but must be considered within the context (e.g. an immediate threat) and ignoring any hyperbole (e.g. acts which are most likely improbable or impossible to happen).[6] The individual, immediate threat as a contextual factor has been upheld in several instances [7,8], while general, overly broad banning

[1] Chaplinsky v. New Hampshire, 315 U.S. 568, 572 (1942)
[2] Cohen v. California, 403 U.S. 15 (1971)
[3] Gooding v. Wilson, 405 U.S. 518 (1972)
[4] Hess v. Indiana, 414 U.S. 105 (1973)
[5] R.A.V. v. City of St. Paul, 505 U.S. 377 (1992)
[6] Watts v. United States, 394 U.S. 705 (1969)
[7] Ovadal v. City of Madison, Wis., 469 F. 3d 625 (Court of Appeals, 7th Circuit 2006)
[8] State v. Clay, 975 SW 2d 121 (MO Supreme Court, 1998)

of offensive speech has been struck down (particularly at college and university campuses) as such speech can be used to communicate general ideas. [9,10,11]

In contrast, laws throughout the EU/Europe, Australia, and Canada (among others) take a much more holistic approach to defining hate speech. Many European countries have laws against denying the Holocaust. Beyond that, in these countries, any speech that calls for genocide, violence, or *more* hatred by one group over another is considered punishable by law. Moreover, many countries prohibit speech that derogates any protected group or incites any form of hostility.

An example of such a holistic view of hateful rhetoric is Germany's law on the "incitement to hatred" ("Volksverhetzung"), which punishes any who satisfy the following criteria:

> Whosoever, in a manner capable of disturbing the public peace:
>
> - Incites hatred against a national, racial, religious group or a group defined by their ethnic origins, against segments of the population or individuals because of their belonging to one of the aforementioned groups or segments of the population or calls for violent or arbitrary measures against them; or
> - Assaults the human dignity of others by insulting, maliciously maligning an aforementioned group, segments of the population or individuals because of their belonging to one of the aforementioned groups or segments of the population, or defaming segments of the population (German Criminal Code, Section 130)[12].

Other, related restrictions follow, including against those who disseminate "such written materials" (i.e. express endorsement), "denies or downplays an act committed under the rule of National Socialism", and "violate[s] the dignity of the victims by approving of, glorifying, or justifying National Socialist rule of arbitrary force".

Hate Speech Studies in Natural Language Processing

In computational studies of hate speech, data from social media sites are collected and labeled by annotators with a definition of hate speech in mind, or a specific set of criteria. And with widespread social media usage, online hate speech definitions have readily emerged with varying resemblance to original theoretical underpinnings. These definitions are operationalized through inconsistent means to real data, with some NLP research specifying this step and

[9] Dambrot v. Central Michigan University, 55 F.3d 1177 (6th Cir., 1995)

[10] Iota Xi Chapter of Sigma Chi Fraternity v. George Mason University., 993 F.2d 386 (4th Cir.,1993)

[11] DeJohn v. Temple University, 537 F.3d 301 (3d. Circuit, 2008)

[12] `https://www.gesetze-im-internet.de/englisch_stgb/englisch_stgb.html#p1246`

others leaving it implicit (i.e. just telling annotators to use the definition without further instructions or criteria).

The NLP community is aware of the drawbacks of this definitional inconsistency: Schmidt and Wiegand (2017) state "the fact that no commonly accepted definition of hate speech exists further exacerbates this situation [of sparse hate speech data]" (p. 8). But in general, NLP studies are not as concerned with the systematic definition of hate speech, and tend to focus on the linguistic content of the speech act. This includes the disambiguation of inherently hateful terms in different contexts, the expression of hate through assertions rather than words, or the referencing of ideological movements through various means (hash-tags, pop cultural references, etc.).

A standard in the NLP literature for studying hate speech is Warner and Hirschberg (2012), which asserts that "hatred against each different group is typically characterized by the use of a small set of high frequency stereotypical words" (p. 19). The authors cite Nockleby's definition of hate speech: "any communication that disparages a person or a group on the basis of some characteristic such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristic" (Nockleby, 2000). They also add further considerations as to the nature of hate speech: endorsements of hateful organizations (such as the KKK) are not hate speech, nor is racial pride. In contrast, they argue that "unnecessary labeling of an individual as belonging to a group often should be categorized as hate speech", as this is a way to "invoke a well known, disparaging stereotype" (p. 20).

A typology of hate speech – designed for racist and sexist speech acts using inversions of privilege as a definition – is proposed by Waseem and Hovy (2016), who list criteria for a Tweet to be considered hate speech, including: "uses a sexist or racial slur"; "seeks to silence a minority"; "criticizes a minority (without a well founded argument)"; and "blatantly misrepresents truth or seeks to distort views on a minority with unfounded claims" (p. 89). This approach, while lacking systematic justification, still represents a step forward in developing a typology of speech acts associated with hate speech. Related work by Olteanu et al. (2018) proposes a coarse typology of "framing" for hate speech:

- *Diagnoses the cause or causes for a problem*
- *Suggests a solution or solutions for a problem*
- *Both diagnoses causes and suggests solutions*

In a work designed to quantitatively differentiate "offensive language" from hate speech, a challenging task given the high level of word-level intersection in these two categories, Davidson et al. (2017) define hate speech as:

> Language that is used to expresses [*sic*] hatred towards a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group (p. 512).

Two categories of speech which are orthogonal to hate speech are "abusive language" (Nobata et al., 2016), which is characterized more by nondescript,

offensive attacks on persons, and "personal attacks" (Wulczyn et al., 2017), which considers flagged comments in Wikipedia discussions.

Research which uses NLP methods to study the content of hate speech includes (Mondal et al., 2017), which uses the "sentence structure" of the articulation of explicit hate (e.g. "I hate black people", p. 4) to identify the targets of hate speech in Twitter. They define hate speech more according to its intent:

> We define hate speech as *any offense motivated, in whole or in a part, by the offender's bias against an aspect of a group of people.*

Other articles which conduct annotation of acts of hate speech offer little to no clarification of what is meant by "hate speech", and largely rely on popular conceptions of "racism", "sexism", or other conventional terms (e.g. Kwok and Wang, 2013; Djuric et al., 2015).

3 Hate-based Rhetoric

As we note in our introduction, *hate-based rhetoric* denotes a class of linguistic acts which partially intersects with hate speech. While more inclusive with respect to contextual and otherwise subjective factors, it also excludes some of the categories of language used in prior NLP studies, including "abusive language" (see Waseem et al., 2017). In formulating and justifying our definition of this class of language, we weave together conceptions of hate speech from the legal and NLP communities (Section 3.1). These sources work together to shape our typology, which we specify in Section 3.2.

3.1 Definition

Our research into hate speech is deeply rooted in trying to understand — and thus prevent or counter — the harm achieved by hate speech and the means by which it is achieved. What marks the difference between offensive language and hate speech (i.e. Davidson et al., 2017)? When can we say that abusive language is motivated by hate or prejudice? What historical contexts, when invoked, constitute an incitement to hatred or violence? In order to address such questions, we look to two dynamics which are at the core of hate speech and prejudice: the assault committed against the dignity of the target, and the intent to commit such assaults by the speaker. Waldron (2012) further explains these concepts:

> A person's dignity is ... their social standing, the fundamentals of basic reputation that entitle them to be treated as equals in the ordinary operations of society ... The publication of hate speech is calculated to undermine this. Its aim is to compromise the dignity of those at whom it is targeted, both in their own eyes and in the eyes of other members of society. It aims to besmirch the basics of their reputation, by associating ascriptive characteristics like ethnicity,

> or race, or religion with conduct or attributes that should disqualify someone from being treated as a member of society in good standing (p. 5).

This holistic definition of hate speech is shared by legal bodies other than the U.S., including Germany. Where the U.S. requires proof of a speech act's relation to the harm of the target, countries like Germany prohibit certain types of rhetoric "not only ... because of their likelihood to lead to harm, but also for their intrinsic content" (Gagliardone et al., 2015, p. 11).

An additional point of emphasis from the German laws on hate speech is the role of historical context in the definition of hate speech. Whereas the U.S. has a restricted view of what constitutes "fighting words", Germany (and other European countries with holistic views of hate speech) prohibit denying/downplaying the Holocaust, as well as other, historically motivated attacks on a previously marginalized or victimized group. In the U.S., many words, stereotypes, and assertions have a particular historical use as a means to insult a particular group, communicate a lesser status about a particular group, or otherwise normalize and extend the power of a dominant group. We can observe this in racial prejudice and other attacks on groups which have a history of being oppressed in their local context.

Thus we are theoretically motivated by two elements of German laws in defining hate speech: its holistic view of the ability of language alone to wound and dehumanize, and its perspective on the role of historical context. And while the German law is perhaps the most famous, any number of other countries' laws could be used in defining hate speech.

From these combined sources, we summarize the definition of *hate-based rhetoric* as the following:

> *Language that intends to — through rhetorical devices and contextual references – attack the dignity of a group of people, either through an incitement to violence, encouragement of the incitement to violence, or the incitement to hatred.*

Definitions of the various types of hate-based rhetoric named above, and the reasoning behind them, are given in the next section.

3.2 Typology

The sources which we use for our definition are also useful in determining the categories which meaningfully delineate the types of hate-based rhetoric. In this section we introduce and justify each of the four dimensions of hate-based rhetoric:

- Hate-based rhetoric: A document can be (1) Not-hateful, (2) Incitement to hatred/Call to Violence; and/or (3) Assault on Human Dignity.
- Vulgarity/Offensive Language: A document can use offensive or abusive language which may or may not be one of the above hate-based categories

- Targeted Group: The type of targeted group
- Implicit/Explicit: Whether the rhetoric is direct and explicit, or it is veiled and reliant on external information to accomplish its objective.

Incitement to Hatred and Incitement to Violence

The question of how to partition the types of hate-based rhetoric is critical. The concept of scale or severity in hate speech has only recently been discussed systematically. Obviously, some hate speech is worse than others, though this was an informal fact until recently. For example, Olteanu et al. (2018) develops a typology with four dimensions of hate speech: stance, target, severity, and framing. The three levels of severity here are: "promotes violence", "intimidates", and "offends or discriminates" (p. 5). A similar hierarchy of hate speech comes from Facebook's "Community Standards", which proposes the following three "tiers": violent and dehumanizing speech; statements of inferiority; and calls for exclusion or segregation[13].

A recent publication on online hate speech by The United Nations Educational, Scientific and Cultural Organization (UNESCO) conceptualizes hate speech as being one of two categories:

I "Expressions that advocate incitement to harm (particularly, discrimination, hostility or violence) based upon the target's being identified with a certain social or demographic group."

II A broader category, one including "expressions that foster a climate of prejudice and intolerance on the assumption that this may fuel targeted discrimination, hostility and violent acts" (Gagliardone et al., 2015, p. 10).

This distinction — between what we might call "incitement to harm/violence", including both statements which *advocate* such incitement and those which actually perform it, and *incitement to hatred* — are echoed in the section of the German hate speech law we cited in the above section. Buyse (2014) also distinguishes the two from a legal perspective, discussing the potentially causal relationship between hate speech and incitement to violence. In our view, therefore, the most natural categorization of hate speech is along these lines: there is language which calls for (or endorses) violence, aggression, exclusion, or segregation of a group of people (or an individual by virtue of their group identity), and there is language which "foster[s] a climate of prejudice and intolerance" through dehumanization or other forms of assaults on human dignity.

Vulgarity and Offensive/Abusive Language

Speech which "incites to violence" is relatively clear, in both the literature and examples from content analyses. The other category, *incitement to hatred*, is less

[13] `https://www.facebook.com/communitystandards/objectionable_content/hate_speech`

clear, especially in the context of existing work in NLP which targets offensive language, abusive language, and incivility. From the above discussion of human dignity given by Waldron (2012), we distinguish the *incitement to hatred* from these other forms of undesirable language by the perceived intent of the speaker to dehumanize, disempower, or subjugate a group (or an individual by virtue of group identity).

This distinction applies most directly to the uses of hateful slurs, such as the n-slur or the c-slur. By our definition, in order for the use of a hateful slur to be *incitement to hatred*, there has to be intent on the part of the speaker which satisfies the above criteria. Therefore, casual uses of these terms (e.g. an insult to a friend, where the group referenced by the slur is not involved) are offensive and worthy of flagging, but not hate-based rhetoric.

Targets of Hate-based Rhetoric

As detailed by Warner and Hirschberg (2012) (who conceptualize hate speech as the use of well-known stereotypes), the *incitement to hatred* varies in form according to the targeted group:

> We ... sub-divide such speech by stereotype, and we can distinguish one form of hate speech from another by identifying the stereotype in the text. Each stereotype has a language all its own, with one-word epithets, phrases, concepts, metaphors and juxtapositions that convey hateful intent. ... Given this, we find that creating a language model for each stereotype is a necessary prerequisite for building a model for all hate speech (p. 21).

Other research has been attentive to the need to label for the targeted group, including Olteanu et al. (2018) and Mondal et al. (2017). We echo the need for including such a category in our hate-based rhetoric typology. The only distinction we make is that we specify the *type* of group named (i.e. religious, political, ethnic/racial, gender, etc.), rather than the group itself. This was done from the simple fact of simplicity: too many groups are named in hate speech to create a generalizable typology for all of them. Instead, we hypothesize that the rhetorical structures of language targeting an ethnic/racial (or other) group will be similar to each other.

Framing: Implicit or Explicit

Lastly, we have independently adopted an aspect of Waseem et al. (2017)'s typology for the sub-tasks associated with the study of abusive language, in that we ask annotators to label each instance as "explicit" or "implicit" hate, which refers to the type of rhetorical device used to express the semantic entity of the sentence or document.

4 Coding Manual/Procedure

In sections below, we instruct annotators on how to apply our definition and typology to real data, including discussions of examples. **Please note that all the below examples have been taken from comments made on YouTube.com and white-supremacist social media accounts and therefore contain hateful rhetoric and abusive language.**

Hate-based Rhetoric

A document can be **CV** (a "Call for Violence"), **HD** (an "Assault on Human Dignity"), **HD** and **CV**, or **NH** ("Not Hateful"). If none apply, the document is to be considered **NH** ("Not Hateful")

Calls for violence (CV)

Calls for violence include any verbalization or promotion of messages which advocate or endorse aggression towards a given person or group on account of their status as member of a given sub-population. This aggression can take the form of violence, genocide, exclusion, and segregation.

Threats which do not name the target's group membership as cause for the threat are **not** hate speech under our definition. Such instances include individual attacks (insults or threats) and group attacks which do not leverage some form of hatred. The below text is such an example of the former:

> paul ryan is a traitor and its too bad we do not hang traitors anymore because he would be just one in a long line

This would be coded as **NH** and **VO**, as it contains a threat which does not target group identity.

A useful set of instructions in coding potentially **CV** documents is given by Benesch (2012):

> Was the speech understood by the audience as a call to violence? Inflammatory speech is often expressed in elliptical, indirect language, which can be variously interpreted. For this analysis, the only relevant meaning is the way in which the speech was understood by the audience most likely to react, at the time when it was made or dessiminated (p. 4).

An example of a document coded as **CV**:

> imagine you were born and living in kongo, or others african stases. i doubt you would had the rights and the lifestyle you fuckers have in usa. stop this racial escalation right now. you dont want the white people become racist for real, dont you? you will regret those days where you were allowed to live free. dont push it too much. you are adviced.

The above text (the last four sentences) endorses and threatens violence against people on account of their race and/or nationality. The above example would also be coded as **HD** (see below), as it implies the inferiority of other nations and cultures.

Assaults on human dignity (HD)

A document should be labeled as **HD** if it assaults the dignity of group by: asserting or implying the inferiority of a given group by virtue of intelligence, genetics, or other human capacity or quality; degrading a group, by comparison to subhuman entity or the use of hateful slurs in a manner intended to cause harm; the incitement of hatred through the use of a harmful group stereotype, historical or political reference, or by some other contextual means, where the intent of the speaker can be confidently assessed.

In the evaluation of slurs against group identity (race, ethnicity, religion, nationality, ideology, gender, sexual orientation, etc.), we define such instances as "hate-based" if they are used in a manner intended to wound; this naturally excludes the casual or colloquial use of hate slurs. As an example, the adaptation of the N-slur (replacing the "-er" with "-a") often implies colloquial usage. Words such as "bitch" and "dick" are to be considered hateful if they are used in a way which dehumanizes the respective, targeted populations.

An example of a **HD** document, which uses a word viewed as inherently hateful/degrading that is not colloquial:

> We grew up in the 50s saying [N-slur], spic, wop, pole-lock, making ethiopian skinny jokes, we joked and laughed at all races and cultures, including ours. hate what the left has done with pc.

Language which dehumanizes targeted persons/groups will also be labeled as **HD**. In coding dehumanizing rhetoric, we refer coders to Haslam (2006), who developed a model for two forms of dehumanization. In *mechanistic* forms, humans are denied characteristics that are "uniquely human" (p. 252). Depriving the other from such traits is considered downward, animalistic comparison. Put another way, the target has been denied the traits that would separate them from animals. An explicit example of such dehumanizing speech:

> you sound so stupid like what is your purpose in life?? dont quit your day job buddy. these youtube videos from you fake black people ruin a lot of what you black stupid traitor monkeys stand for.

In another form of dehumanization as categorized by Haslam (2006), the target may be denied qualities related to human nature. These characteristics are traits that may not be unique to humans, but define them. These traits will "represent the concept's 'core' [but] may not the same ones that distinguish us from other species" (p. 256). When these traits are denied from the target, this is considered upward, mechanistic dehumanization. The result of denial is often perceiving the target as cold, robotic, and lacking deep-seated core values and characteristics.

Documents which invoke cultural, political, or historical context in order to voice negative sentiment/degradation toward a particular sub-population, empower hateful ideology (hate groups), or reduce the power of marginalized groups, are to be considered **HD** as well. This would include messages which indicate support for white supremacy (e.g. advocating for segregated societies/apartheid), those which make negative assertions and/or implications about the rights of certain groups (e.g. "Immigrants in this country need to go back to their country"), and those that reduce the power/agency of particular segments of the population.

An example of reducing the power of a segment of the population through perpetuation of human degradation:

> fun fact: black people stole other black people and sold them to slavery, white people just happened to buy them i am not racist, and putting a comment out like this is oddly hard.

An example of elevating the status/power of a hateful group/ideology (anti-Semitism):

> even if all the jews in europe suddenly moved to another continent, there would still be white traitor political parties who approve of massive immigration. they are the head - not the jew. cut off the head and the whole treasonous government will die.

HB_derogatory: Vulgarity/Offensive Language directed at an individual (VO)

As we alluded to in the above sections, some documents will contain derogatory/offensive language or personal attacks, but may or may not be classified as *hate-based rhetoric*. Strictly speaking, derogatory/offensive language (**VO**) is only violating human dignity (**HD**) if the offensive language targets a group's characteristics as a negative/degrading comparison that asserts or implies a group to be less than others. Similarly, attacks or insults (**VO**) directed at individuals are only calls for violence (**CV**) when they are justified by the subject's membership in a group or segment of the population.

Examples of vulgarity that are not hate (**NH**):

> And it's gotta be some ghetto fool :/

> All these bitches want a baby, I don't want no children.

> #ThingsNotToDoOnAFirstDate give her all da dick U gotta half stroke da pussy bc if u dump dick on dat bitch she gon become extremely anoyin

Examples that are considered an Assault on human dignity/perpetuation of human degradation (**HD**) that contain vulgar, derogatory, and offensive language (**VO**):

Anti-Woman:

> Load of fucking cunts at the bbc dumb bitches who have never dated anyone above an iq of 85 or talked to an intelligent man i pity them then again thats why they work at buzzfeed now now you cunts keep it classy you are all pretty.

Anti-Black:

> hey [N-slur] north africa isnt black just thought you should know by the way youre clearly a racist yourself so why youre trying to shame anyone is beyond me

If you are unsure that a word is being used as a derogatory slur directed in a hateful, degrading way towards a segment of the population, feel free to consult `hatebase.org` or `urbandictionary.com`.

Targeted Populations

This column should be used to identify which types of groups are targeted by the hateful rhetoric. Select all descriptors which apply.

- **RAE**: Race or ethnicity (includes anti-asian, anti-latino, anti-black, anti-arab, anti-semitism etc.)
- **NAT**: Nationality/regionalism (includes general xenophobia and targets against countries/regions)
- **GEN**: Gender (anti-woman, anti-man, anti-trans etc.)
- **REL**: Religion/spiritual beliefs (anti-muslim, anti-christian, etc.)
- **SXO**: Sexual Orientation
- **IDL**: Ideology (conservative/liberal/leftist/right-wing)
- **POL**: Political identification. Includes any reference to membership in a political organization (Democratic/Republican/ etc.).
- **MPH**: Mental/physical health status, physical disability.

Framing: Implicit or Explicit

A document in our typology of *hate-based rhetoric* can be either explicit or implicit with regard to the communicated message. Waseem et al. (2017) articulates this as:

> This is roughly analogous to the distinction in linguistics and semiotics between denotation, the literal meaning of a term or symbol, and connotation, its sociocultural associations (p. 2).

Waseem et al. (2017) also explains that explicit rhetoric is "unambiguous in its potential" to be hateful, whereas implicit rhetoric "does not immediately imply or denote" hatred (p. 2). This rhetoric is typically characterized by sarcasm, irony, metaphor, and the lack of explicitly hateful terms.

An example of implicit rhetoric:

> just because you were born in america, does not make you american.
> just as if a white man is born in japan, he is not japanese.

As this text carries the implication that there are ethnic/national requirements for being considered "American", and is therefore an act of exclusion; however, we have to assume this communicative intent on behalf of the speaker. This speech is xenophobic (**NAT**).

If the rhetoric is made explicit and there is no ambiguity/subliminal messaging, assign **EX** (Explicit) to this column.

Notes

Is there anything important to note about the document? Is it in a different language? Does it seem in any way like it should be excluded (e.g. because it is fake, because it has been repeated multiple times, etc.) If so, add a note describing these issues in the 'note' column.

In addition, make notes about difficult cases and difficulties you had in coding, as we will use these contributions to refine and improve later versions of this manual.

5 Examples

We have gathered a more comprehensive set of examples for the illustration of the application of our definition of *hate-based rhetoric*. **Please note that all the below examples have been taken from comments made on YouTube.com and white-supremacist social media accounts and therefore contain hateful rhetoric and abusive language.**

Gab examples				
Text	Hate-Based Rhetoric	Derogatory	Target	Framing
The Black Panther movie is a blatant anti-white ZioJesuit cultural marxist film aimed at further anti-white Christian theft, replacement, and genocide. Send them all back to their native Africa like good ole Abe had the foresight to do and let them find equality and social justice among their own kindred and creeds.	HD, CV	VO	SXO, RAE	EX
Labours New LGBT+ Adviser: All White People Are Racist ... Yeah! We don't want our white Christian nations invaded and destroyed by our non-white and non-Christian enemies using nasty homosexual degenerate sockpuppets	HD	VO	RAE, REL, SXO	EX
So, you just hate all white people who don't want to race mix their people out of existence or allow enemy people of color to invade their nations in order to rob, and destroy their advanced race, creed, wealth, and culture.	HD		RAE	EX
Your existence, privilege, and success as a race are built on the backs, blood, and death of people of color. BS! Whites built their own nations with their own blood, backs, sweat, tears, brains, and strength in the name of Christ. Africans were used by Jew/Jesuit slavers to control all the white slaves who were the actual majority.	HD		RAE	EX
The problem is that we're not speaking a language the #IslamoFascistsunderstand.We speak of due process and rights. They only understand force and death.If we want to send a real message, it can't be verbal. It must be bloodshed. It must be a public execution, followed by the promise of more.	CV		REL	EX
Ok so apparently they will be sent back to shitholes.Like I care? Nope.	HD		NAT	IM

So a wetback can sneak into our country six times, play with a gun, kill a woman, and mostly get away with it.You can literally get away with murder... So long as you're a poor immigrant. #JusticeForKate	HD	VO	RAE	EX
If they were all really just illegals, refugees or various forms of so-called legal migrants we would not now really have 40+ million non-white invaders and their anchor kids now bankrupting, robbing, replacing & killing whitey along with our cops. We would not be watching all the anti-white non-sense on Jew/Jesuit Talmudic media. An enemy has done this.	HD		RAE, NAT	EX
Drunk Illegal Crashes into Ambulance; Kills3-Year-Old...Where's the lame-stream Jew/Jesuit news media?	HD		RAE	IM
Seriously, Italy? You chose a Nigerian named Chike Iwobi to be the face of Italian nationalism? God help us.	HD		NAT	IM
Each of our classical works now has a forward from some Jew telling us what to think about it. They must be expelled, by force if need be.	CV		RAE	EX
Africans will always be savages. No matter the period of history, education they receive or country they live in.	HD		RAE, NAT	EX

Examples from Alt-right YouTube Comments				
Text	Hate-Based Rhetoric	Derogatory	Target	Framing
i hate milo because hes a degenerate faggot [N-slur] fucker. he should be deported and sent to a muzzie country and get thrown off a roof	HD, CV	VO	SXO, RAE	EX
in fact, both muslim and non-muslim middle easterner and africans alike are turning europe into a shithole, and fast - while displacing native europeans all at once.	HD		NAT	EX
man nancy pelosi is a fucking cunt. i cant imagine how someone with so much power can be either this delusional or fucking corrupt.	NH	VO	GEN	EX
the black people seem determined to march back to the plantation	HD		RAE	IM

6 Future work

The most significant short-coming of our typology and coding procedure is the lack of a theory of speech acts, similar to that provided by Olteanu et al. (2018) but possessing a more general character. In our future work in developing and refining this manual, a primary goal is to use real-data examples to add to and embellish our typology of speech acts which are used in *hate-based rhetoric*; for now, we are limited to the division between implicit and explicit rhetoric.

Another limitation of our typology is the accounting for the various ways in which slurs are employed: as insults, as means of dehumanization, as benign discussions of the slurs themselves, etc. Defining these types of language is an active and controversial research area, one we hope to contribute to by considering their communicative intent.

An ongoing project, which we have already begun, is the application of our manual to real data sets. Currently, we have annotated the twitter data set made available by Davidson et al. (2017), and a collection posts from the "Gab" social networking service (forthcoming).

References

Anderson, A. A., Brossard, D., Scheufele, D. A., Xenos, M. A., and Ladwig, P. (2014). The "nasty effect:" online incivility and risk perceptions of emerging technologies. *Journal of Computer-Mediated Communication*, 19(3):373–387.

Benesch, S. (2012). Dangerous speech: A proposal to prevent group violence. *Voices That Poison: Dangerous Speech Project*.

Buyse, A. (2014). Words of violence: Fear speech, or how violent conflict escalation relates to the freedom of expression. *Human Rights Quarterly*, 36:779.

Davidson, T., Warmsley, D., Macy, M., and Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *arXiv preprint arXiv:1703.04009*.

Djuric, N., Zhou, J., Morris, R., Grbovic, M., Radosavljevic, V., and Bhamidipati, N. (2015). Hate speech detection with comment embeddings. In *Proceedings of the 24th international conference on world wide web*, pages 29–30. ACM.

Gagliardone, I., Gal, D., Alves, T., and Martinez, G. (2015). *Countering online hate speech*. Unesco Publishing.

Haslam, N. (2006). Dehumanization: An integrative review. *Personality and social psychology review*, 10(3):252–264.

Kwok, I. and Wang, Y. (2013). Locate the hate: Detecting tweets against blacks. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*. AAAI.

Mondal, M., Silva, L. A., and Benevenuto, F. (2017). A measurement study of hate speech in social media. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media*, pages 85–94. ACM.

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., and Chang, Y. (2016). Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, pages 145–153. International World Wide Web Conferences Steering Committee.

Nockleby, J. T. (2000). Hate speech. *Encyclopedia of the American constitution*, 3:1277–79.

Olteanu, A., Castillo, C., Boy, J., and Varshney, K. R. (2018). The effect of extremist violence on hateful speech online. *arXiv preprint arXiv:1804.05704*.

Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., and Wojatzki, M. (2017). Measuring the reliability of hate speech annotations: The case of the european refugee crisis. *arXiv preprint arXiv:1701.08118*.

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

Sellars, A. (2016). Defining hate speech. *Berkman Klein Center Research Publication*, 2016(20).

Waldron, J. (2012). *The harm in hate speech*. Harvard University Press.

Warner, W. and Hirschberg, J. (2012). Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*, pages 19–26. Association for Computational Linguistics.

Waseem, Z., Davidson, T., Warmsley, D., and Weber, I. (2017). Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84.

Waseem, Z. and Hovy, D. (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.

Weinstein, J. (2018). *Hate speech, pornography, and radical attacks on free speech doctrine*. Routledge.

Wulczyn, E., Thain, N., and Dixon, L. (2017). Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1391–1399. International World Wide Web Conferences Steering Committee.