# The Gab Hate Corpus: A collection of 27k posts annotated for hate speech

Brendan Kennedy[1], Mohammad Atari[2], Aida Mostafazadeh Davani[1], Leigh Yeh[1], Ali Omrani[1], Yehsong Kim[2], Kris Koombs[3], Shreya Havaldar[1], Gwenyth Portillo-Wightman[2], Elaine Gonzalez[2], Joe Hoover[2], Aida Azatian[†2], Alyzeh Hussain[†2], Austin Lara[†2], Gabriel Olmos[†2], Adam Omary[†2], Christina Park[†2], Clarisa Wang[†2], Xin Wang[†2], Yong Zhang[†2], and Morteza Dehghani[1,2]

[1]Department of Computer Science, University of Southern California

[2]Department of Psychology, University of Southern California

[3]Department of Political Science, University of Southern California

## Author Note

[†] authors contributed equally. Correspondence regarding this article should be addressed to Brendan Kennedy, btkenned@usc.edu, 362 S. McClintock Ave, Los Angeles, CA 90089-161. This research was sponsored by NSF CAREER BCS-1846531 to MD.

Abstract

The growing prominence of online hate speech is a threat to a safe and just society. This endangering phenomenon requires collaboration across the sciences in order to generate evidence-based knowledge of, and policies for, the dissemination of hatred in online spaces. To foster such collaborations, here we present the Gab Hate Corpus (GHC), consisting of 27,665 posts from the social network service `gab.ai`, each annotated by a minimum of three trained annotators. Annotators were trained to label posts according to a coding typology derived from a synthesis of hate speech definitions across legal, computational, psychological, and sociological research. We detail the development of the corpus, describe the resulting distributions of hate-based rhetoric, target group, and rhetorical framing labels, and establish baseline classification performance for each using standard natural language processing methods. The GHC, which is the largest theoretically-justified, annotated corpus of hate speech to date, provides opportunities for training and evaluating hate speech classifiers and for scientific inquiries into the linguistic and network components of hate speech.

*Keywords:* Hate speech, Theory-driven text analysis, Natural language processing, Social Media, Open Data, Text Annotation

# The Gab Hate Corpus: A collection of 27k posts annotated for hate speech

On the morning of October 27th, 2018, an anti-Semite walked into a synagogue in Pittsburgh, Pennsylvania, and opened fire, killing eleven and wounding six (Bradbury, 2018). The gruesome act of violence quickly gained attention due to the prejudiced motivations of the shooter. The attack was against the Jewish people, motivated by an utterly unfounded, intense hatred, and was documented by the perpetrator on his social media account on "Gab" shortly before the act (Roose, 2018). `gab.ai`, an online social network that claims to be devoted to the preservation of free speech, has become inhabited by deplatformed white nationalists, neo-Nazis, and other hate-mongering ideologues (Grey Ellis, 2016), and is one of several online communities supporting hateful and abusive rhetoric (Matsakis, 2018). Understanding and mitigating the spread of hatred, within online platforms and extending to associated behavioral outcomes, is a clear challenge for the modern world.

Political scientists, psychologists, and sociologists have studied hatred for decades, theorizing the ways in which prejudice forms (Allport, Clark, & Pettigrew, 1954) and manifests in violence (Müller & Schwarz, 2019), investigating the psychology of hate groups (Glaser, Dixit, & Green, 2002), and documenting the impact of hatred on victims and the marginalized (e.g., Herek, Gillis, & Cogan, 1999; McDevitt, Balboni, Garcia, & Gu, 2001). Concurrently, legal scholars and public policy experts have debated and implemented strategies for combating hate crime and hate speech, perpetually engaged in the debate of free speech or hate speech censorship (e.g., J. W. Howard, 2019; Sellars, 2016). How online hate speech fits into these scientific and legal precedents is still being negotiated. The United Nations Educational, Scientific and Cultural Organization (UNESCO; Gagliardone, Gal, Alves, & Martinez, 2015) delivered a program for countering online hate speech, outlining its particular challenges. Technology companies including Facebook (*Community Standards: 12 Hate Speech*, 2020), Twitter (*Hateful conduct policy*, 2020), and Google (*Hate speech policy*, 2020) have posted official policies on hate speech and abusive language

that sometimes go beyond legal restrictions. The priority of these policies is to limit human exposure to hateful content, and to facilitate the effective removal of harmful content from online platforms; to this end, natural language processing (NLP) researchers have devoted significant resources to developing detection algorithms for hate speech, abusive language, and offensive language (e.g., Davidson, Warmsley, Macy, & Weber, 2017; Warner & Hirschberg, 2012; Waseem & Hovy, 2016).

The emergence of online hate speech and the growing problem of online radicalization are part of the more general phenomena of outgroup hatred and prejudice. It thus might be expected that perspectives from social scientists on these phenomena, which are undeniably social, psychological, and political, can inform and evaluate policies and detection strategies. While interdisciplinary work has begun to occur in hate speech detection research, the current level of communication and collaboration between social scientists, computer scientists, and legal and public policy experts is unsatisfactory. Definitions of hate speech in computational work are mostly ad-hoc and atheoretical, potentially influenced by the lack of legal consensus in defining hatred and hate speech. Research on hatred and prejudice in the social sciences is inhibited by the relative inaccessibility of hate crimes and hate speech. Both are rare events that cannot be studied in laboratory settings, individuals are likely to self-censor when reporting their attitudes towards hate crime or hate speech, and engagement with hateful groups or ideas. Consequently, definitions of hatred and prejudice in the social sciences are rarely operationalized in quantitative analyses. Importantly, this precludes the possibility of naturalistic observations of hateful behaviors and language, which Rozin (2001) and others have identified as critical for the "...description of universal or contingent invariances" (p. 3) in socio-psychological phenomena.

In this work, we (a) review and synthesize prior work in multiple disciplines to develop a theoretically-justified typology and coding guide (see Appendix A), (b) rigorously train a cohort of undergraduate research assistants to accurately identify hate-based

rhetoric based on the developed coding guide (see Methods), (c) annotate 27,665 posts from the social network platform Gab by a minimum of three trained annotators per post (see Results), (d) run baseline as well as state-of-the-art machine learning models to classify the entirety of the Gab corpus (see Results), and (e) publicly release this expert-annotated large-scale dataset for usage in natural language processing research as well as computational linguistic studies on psychology of hate speech (see Discussion).

**Synthesizing Perspectives on Hate Speech**

Historically, hate groups and hate crime have primarily been discussed within the legal domain, centering on discussions of whether acts of hate, including violence, intimidation, and defamation, are protected as free speech or ought to be criminalized (J. W. Howard, 2019; Sellars, 2016). The term "hate speech" was coined fairly recently in reference to a particular set of offenses under the law: Matsuda (1989) argued that "the active dissemination of racist propaganda means that citizens are denied personal security and liberty as they go about their daily lives" (p. 2321). Online hate speech, which possesses marked differences from offline hate speech in terms of legal precedent (Gagliardone et al., 2015), has begun to be policed not by countries, but by technology companies, in part due to inconsistent free speech protections with regards to hateful or offensive language. In the U.S., which prides itself on its freedoms, particularly those of speech (J. W. Howard, 2019), the "lewd and obscene, the profane, the libelous and the insulting or 'fighting' words" (*Chaplinsky v. New Hampshire*, 1942) are prohibited, but what is commonly referred to as hate speech is viewed as the expression of a political idea (*RAV v. St. Paul*, 1992). In contrast, Germany, Australia, the Netherlands, and others (see J. W. Howard, 2019) protect fewer classes of prejudicial expression. For example, Germany's "Volksverhetzung" ("incitement to hatred") law prohibits "Assaults [on] the human dignity of others", including "den[ying] or downplay[ing] an act committed under the rule of National Socialism" and "violat[ing] the dignity of the victims by approving of, glorifying, or justifying National Socialist rule of arbitrary force" (*German Criminal Code*,

1998, Sec. 130).

The latter perspective identifies hate speech according to its motivations and makes explicit the cultural and societal contexts that inform its recognition by human judges — e.g., condemning Holocaust denial recognizes the Holocaust and the intentions of those trying to fabricate history. This focus on motivations and context is more appropriate for those attempting to quantitatively study the socio-psychological components of organized hatred and prejudice that are observed in hate speech. Pragmatically, it is also more aligned with efforts to counter hate speech, rather than simply detecting and censoring it (Gagliardone et al., 2015; Waldron, 2012).

Other operationalizations of hate speech come from natural language processing (NLP) researchers, who have developed algorithmic approaches to detect hate speech based on a mix of manually coded examples (e.g., Schmidt & Wiegand, 2017; Warner & Hirschberg, 2012; Waseem & Hovy, 2016) and keyword-based filtering (e.g., Davidson et al., 2017; Olteanu, Castillo, Boy, & Varshney, 2018). Definitions used in manually coding display task-specific and data-specific features, for example the enumeration of specific stereotypes (e.g., Warner & Hirschberg, 2012) and offensive violations (e.g., Waseem & Hovy, 2016). Related categories of language similar to hate speech are also considered in NLP research, including abusive language (e.g., Nobata, Tetreault, Thomas, Mehdad, & Chang, 2016), incivility (e.g., Anderson, Brossard, Scheufele, Xenos, & Ladwig, 2014), hateful stereotypes (e.g., Warner & Hirschberg, 2012), offensive language (e.g., Davidson et al., 2017), and personal attacks (e.g., Wulczyn, Thain, & Dixon, 2017). Like Germany, the Netherlands, Canada, Australia, and others, some NLP research uses definitions that take into account societal and cultural context. For example, Warner and Hirschberg (2012) consider hate speech to be "harmful stereotypes", which require a level of cultural knowledge or familiarity. Similarly, Waseem and Hovy (2016) identified one type of offensive language as the expression of support for harmful ideologies on social media, which is protected free speech in the U.S. but can be unlawful in countries like Germany,

depending on the ideology being supported. However, much of the research has focused on the identification of disallowed behavior on an online platform, rather than identifying violations according to a specific country's hate speech legislation.

While abusive (e.g., insults) or offensive (e.g., swearing) language are defined by form, hate speech is defined by function. It's functional role depends on the relationship between the perpetrator and the target's group; it is designed by the perpetrator 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" (Waldron, 2012, p. 5). The action of hate speech is critical: to attack, assault, or subvert another individual or group's standing. This assault is not random, but calculated to achieve some social utility. Perry (2002) wrote of hate crime, "Bias-motivated crime provides an arena within which white males in particular can reaffirm their place in a complex hierarchy and respond to perceived threats from challengers of the structure — especially immigrants, people of color, women, and homosexuals" (p. 3). Understanding the motivations and conditions for hate-motivated violence must take into account the existing cultural and social context in which it occurs: "Hate crime ... is much more than the act of mean-spirited bigots. It is embedded in the structural and cultural context within which groups interact ... [I]t is a socially situated, dynamic process, involving context and actors, structure, and agency" (Perry, 2002, p. 2).

Thus the definition of hate speech in this work is not strictly legal, nor is it socio-psychological in nature, rather it is a synthesis of what the hate crime and hate speech literature collectively regards as the "intentional verbalization of prejudice against a social group". We refer to this construct as "hate-based rhetoric", which is distinct from any particular definition in the legal or public policy community, though it bears overall similarities to definitions of hate crime in sociology and psychology.

## Mapping Hate-Based Rhetoric to Data

In computational research, the supremacy of supervised machine learning — i.e., statistical learning by example — has made it one of the default paradigms for automating information extraction from large datasets. Recent work in computational social science, in particular the Moral Foundations Twitter Corpus (MFTC; Hoover et al., 2020), have applied the principles of supervised learning to text data, thereby facilitating out-of-sample prediction and analyses of the importance of linguistic features of socio-psychological phenomena. The annotation of hate speech is widely practiced in computational research (e.g., Davidson et al., 2017; de Gibert, Perez, García-Pablos, & Cuadros, 2018; Warner & Hirschberg, 2012) for the purposes of training and evaluating detection algorithms. The present work shares this objective, and more: annotating socio-psychological constructs such as hate speech or moral sentiment additionally requires the operationalization of sometimes ill-defined or under-defined phenomena, encouraging more fine-grained conceptualizations of the construct in question.

For example, typological definitions of hate speech, while sometimes lacking in theoretical justification, can inform on its sub-components, such as the groups targeted or the rhetorical devices used. Of note is the approach of Warner and Hirschberg (2012), which described hate speech according to its variable linguistic forms depending on the group being targeted: "[E]ach stereotype has a language all its own, with one-word epithets, phrases, concepts, metaphors and juxtapositions that convey hateful intent ... Anti-hispanic [*sic*] speech might make reference to border crossing or legal identification. Anti-African American speech often references unemployment or single parent upbringing. And anti-semitic [*sic*] language often refers to money, banking and media" (p. 21).

## A Typology of Hate-Based Rhetoric

Informed by our review of scholarly work in the area of hate speech, we apply the typology-driven methodologies of computational research towards a typology of hate-based

rhetoric. This hierarchical coding typology was used to facilitate a more consistent, informed annotation across annotators, as well as develop more structured information, such as the categorization of target populations (Mondal, Silva, & Benevenuto, 2017) and framing effects (Olteanu et al., 2018; Waseem, Davidson, Warmsley, & Weber, 2017). The full version of the typology used to train annotators, which includes our definitions, explanations of categories, and examples, is included in Appendix A.

Hate-based rhetoric is determined by the extent to which it is dehumanizing, attacking human dignity, derogating, inciting violence, or supporting hateful ideology, such as white supremacy. An important component of hate-based rhetoric is that occurrences of such language are explicitly directed towards a social group. According to our presented typology, documents were coded "assaults on human dignity" (HD) or "calls for violence" (CV) if they satisfied both of these criteria. Specifically, the HD category broadly includes the assertion or implication of inferiority of a given group by virtue of intelligence, genetics, or other human capacity or quality; degrading or dehumanizing 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 the endorsement of a known hate group or ideology. This categorization is specifically supported by legal codes in Germany, which illegalize speech "not only . . . because of their likelihood to lead to harm, but also for their intrinsic content" (Gagliardone et al., 2015, p. 11).

The separation of HD and CV was done according to legal precedent, best summarized by the following two-class specification given by The United Nations Educational, Scientific and Cultural Organization (UNESCO): (a) "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; (b) A broader category 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). Thus language classified with CV was judged to be a particular incitement to violence, which either directly or indirectly called for or otherwise advocated violence against a group or an individual because of their group membership.

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. In addition, phrases such as "I hate my mother-in-law's guts" should not be classified as hate speech as the target is not hated for their group identity. Our full typology and workflow that annotators were to follow is visualized in Figure 1.

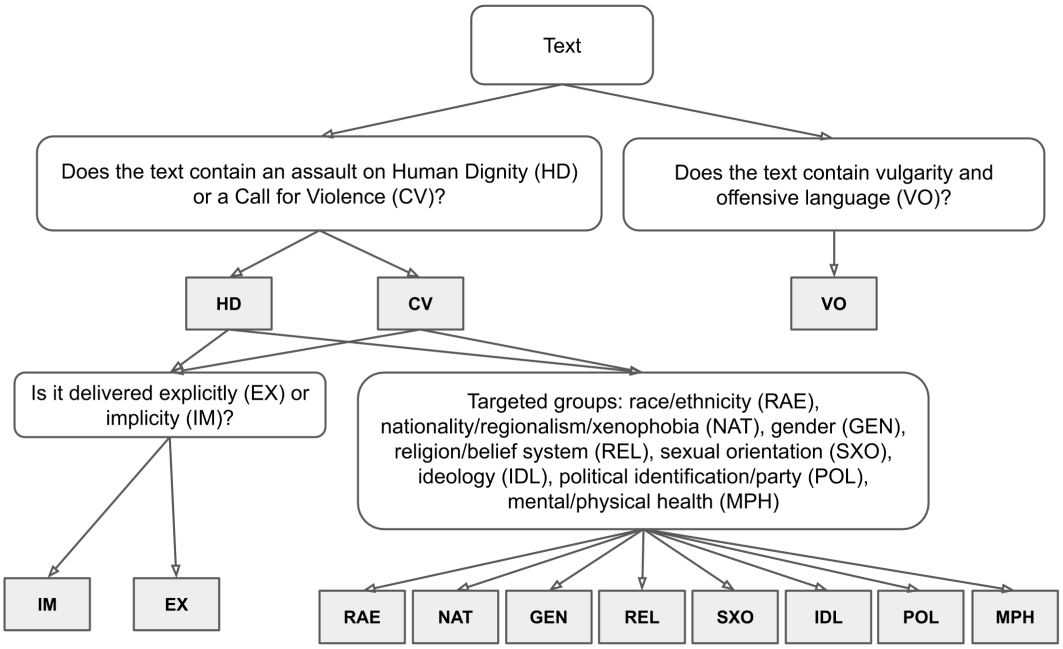


*Figure 1*. Procedure for coding a text according to the "hate-based rhetoric" definition. Primary categories for hate speech are HD and CV, while VO can apply to both hateful and non-hateful texts. Framing (IM/EX) and targeted group apply only to hateful texts, and targeted group labels are non-exclusive

According to the review of the literature and our proposed annotation schema,

independent of whether or not a document is hateful, we annotate documents according to their usage of "Vulgarity and/or Offensive language" (VO). Both innocent and malicious usage of slurs and insults can be VO without being HD or CV, if there is no *group*-level attack (i.e., the attack is against an individual and not on account of their group-level characteristics). Offensive language (i.e., VO) is only violating human dignity (HD) if targeting a group or a group's characteristics. 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. In terms of attacked group identity, we label attacks on nationality/regionalism (e.g., xenophobia), race or ethnicity (e.g., anti-Black), gender (e.g., anti-woman, anti-man, anti-trans), religious or spiritual identity (e.g., anti-Muslim), sexual orientation (e.g., anti-lesbian), ideology (e.g., anti-"leftist"), political identification (e.g., anti-Republican), and mental or physical health status (e.g., ableism). Lastly, we used a single binary dimension in an attempt to annotate texts' "framing" effects. Consistent with Waseem et al. (2017), who introduced the notion of "explicit" versus "implicit" speech as "...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), we code texts for implicit (vs. explicit) framing. Implicit rhetoric is most often an invocation of derogatory beliefs, sentiments, or threats which are accessible through cultural knowledge.

**Protecting Annotators**

We aim to accurately apply grounded definitions of hate speech to text in order to build a comprehensive resource of annotated hate speech. In addition to the lack of theoretical coherence in hate speech coding typologies used in previous computational work, a pressing concern is the potential harm for annotators in producing these valuable data resources.

While no research has yet examined the effects of consistent, daily exposure to hate

speech on human moderators, there is evidence that exposure to online abuse may have negative mental health consequences (Levin, 2017; Ybarra, Mitchell, Wolak, & Finkelhor, 2006). Exposure to hate speech is associated with symptoms of trauma exposure, higher liability assessments of targets, and low self-esteem (Boeckmann & Liew, 2002; Leets, 2002). A number of technology companies have recently been sued by their workers claiming that moderating traumatic content resulted in post-traumatic stress disorder (PTSD; Hern, 2019; Levin, 2017). Within the DSM-5, the premier diagnostic manual for psychological disorders, second-hand exposure to traumatic material can lead to PTSD when exposure is "repeated or extreme," of which content moderation is both (American Psychiatric Association, 2013). PTSD symptoms resulting from indirect trauma exposure through work is common enough that in the literature there are a number of terms to describe these mental health consequences, including "secondary traumatic stress", "compassion fatigue", and "vicarious traumatization" (Ludick & Figley, 2017; May & Wisco, 2016). These consequences have been studied in police officers, first responders, and social workers, among others (Kleim & Westphal, 2011; Perez, Jones, Englert, & Sachau, 2010; Wagaman, Geiger, Shockley, & Segal, 2015), but not yet content moderators. Thus, while no research has directly tested the effects of continuous exposure to hateful or violent content in moderators, such consequences should be investigated further, as most of online hate speech is currently policed by human moderators. Furthermore, if similar patterns of secondary traumatic stress are found in this group, lessons may be learned from secondary trauma prevention and treatment of other trauma-exposed professionals (Bell, Kulkarni, & Dalton, 2003; Kaplan, Bergman, Christopher, Bowen, & Hunsinger, 2017; Kleim & Westphal, 2011).

## Methods

### Corpus construction

Sampling documents for hate speech annotation is difficult due to sparsity and are consequently non-representative (Wiegand, Ruppenhofer, & Kleinbauer, 2019). Since randomly sampling posts from venues like Twitter results in very few hate speech examples, keywords and topic-based sampling strategies are used to collect data. This biases hate speech datasets to the given keywords and topics. Such non-representativeness is most problematic when considering the generalizable quality of predictors: models that are built on a narrow slice of data (e.g., containing *only* documents with group names like "black", "women", and "Muslim") are unlikely to generate accurate predictions for documents in a wider, more diverse sample of language (Wiegand et al., 2019). Further, models might be biased to detect hate speech transgressions against certain groups, but miss those against groups not represented by keywords or topic-based filtering (Dixon, Li, Sorensen, Thain, & Vasserman, 2018).

Resolving the data quality issues that come from biased sampling is an open issue. In our research we select a non-representative online social community and sample posts totally randomly, providing a corpus more representative of the language of hate speech (though potentially atypical in other ways, given the irregularity of a community that frequently posts hate speech). An independently developed study also takes this approach to gathering hate speech data, considering the online platform of "Stormfront", which is populated by white nationalists (de Gibert et al., 2018). We downloaded Gab posts from the public dump of the data by Pushshift.io[1] (Gaffney, 2018) and randomly sampled ∼28,000 posts for annotation. Posts were sampled only if they met a reasonable threshold for textual content (at least three non-hyperlink tokens).

[1] `https://files.pushshift.io/gab/`

### Annotation process

Annotators were undergraduate research assistants (RAs) trained by first reading the typology and coding manual and then passing a test of about thirty messages that had been previously annotated and agreed upon in terms of their hate-based rhetoric labels. All materials are available in Appendix A. RAs performed annotation via a secure online platform, which provided them with the option to halt annotation at any time.

Annotators were provided with a written guide to prevent secondary trauma, which encouraged annotators to pay attention to signs of hyperarousal, attend to changes in cognition, take breaks, and avoid picturing traumatic situations. It also encouraged annotators to contact researchers if they were experiencing symptoms of PTSD, which were also listed on the guide. This guide attempted to normalize feeling negatively impacted by the work, provide trauma- specific education, help monitor for signs of traumatic stress, and provide a mechanism of support as preventative measures against secondary traumatic stress (Bell et al., 2003). While this measure was put into place to reduce the risk of secondary traumatic stress, several annotators dropped out of the study due to the burden associated with annotating hate speech. Future researchers may consider implementation of other preventative and treatment interventions for secondary traumatic stress including repeated assessment of vulnerability factors, identification of at-risk groups within annotators for traumatic stress, continuous self-care and supervision, and mindfulness training (Kaplan et al., 2017; Kleim & Westphal, 2011).

The number of posts annotated per annotator ($n$ = 18, $M$ = 5,109, $Mdn$ = 4,044) ranged from 288 to 13543. Posts were included that were annotated by at least three annotators.

### Inter-Annotator Agreement

Annotating hate speech has been documented to result in high levels of annotator disagreement (e.g., Ross et al., 2017), attributed to a combination of factors, including

annotator differences in understanding of the definition of hate speech, interpretations of the annotated texts, or evaluating harms done to certain groups (i.e., inconsistent application of the hate speech definition to different social groups; see Mostafazadeh Davani, Atari, Kennedy, Havaldar, and Dehghani (2020)). To evaluate the overall levels of inter-annotator agreement, we computed Fleiss's kappa for multiple annotators (Fleiss, 1971) as well as the Prevalence-Adjusted and Bias-Adjusted Kappa (PABAK; Byrt, Bishop, & Carlin, 1993). PABAK is appropriate with imbalanced labels, which are present in this corpus, as kappa is known to underestimate annotator agreement in such cases.

In Table 1, we list Fleiss's kappa and PABAK for the three top-level labels in our typology (HD, CV, VO), where kappas are computed over the entire corpus ($N$ = 27,655). We also provide the binary label distribution (percent positive) for each using a majority-vote of annotators (ties broken towards positive). In Table 2, we provide Fleiss's kappa and PABAK for only annotations that have either HD or CV as positive. We provide the binary label distribution for each target and framing label overall, as well as for positive HD documents (by majority vote) and positive CV documents (by majority vote).

Table 1
*Inter-annotator agreement kappas and binary label distribution (percent positive) of majority vote-aggregation for assaults on human dignity (HD), calls to violence (CV), and vulgar or offensive language (VO).*

| | HD | CV | VO |
|---|---|---|---|
| Fleiss | 0.23 | 0.28 | 0.30 |
| PABAK | 0.67 | 0.97 | 0.79 |
| Positive % | 8.5% | 0.6% | 6.3% |

**Distribution of Aggregated Labels**

Aggregating each annotator's labels into one label set per document is left to the user of the corpus. In the present work, we aggregate by majority-vote for descriptive purposes, a standard procedure when dealing with multiple annotations (e.g. Garten,

Table 2
*Inter-annotator agreement kappas and binary label distributions for target population and framing labels. Kappas are computed from the set of annotations where either HD or CV is labeled as positive. Target labels denote hate-based rhetoric on basis of religion (REL), race/ethnicity (RAE), sexual orientation (SXO), ideology (IDL), nationality (NAT), political affiliation (POL), and mental or physical health status (MPH). Framing labels denote hate-based rhetoric which is explicit (EX) or implicit (IM).*

| | REL | RAE | SXO | IDL | NAT | POL | MPH | EX | IM |
|---|---|---|---|---|---|---|---|---|---|
| Fleiss | 0.65 | 0.62 | 0.70 | 0.42 | 0.43 | 0.47 | 0.32 | 0.16 | 0.22 |
| PABAK | 0.76 | 0.67 | 0.92 | 0.67 | 0.72 | 0.66 | 0.92 | 0.20 | 0.37 |
| Overall (27665) | 3.1% | 3.5% | 0.7% | 1.6% | 1.5% | 2.3% | 0.2% | 5.9% | 1.8% |
| Majority HD (2349) | 21.9% | 29.4% | 5.9% | 10.5% | 9.6% | 14.5% | 1.6% | 51.0% | 14.0% |
| Majority CV (155) | 18.7% | 13.5% | 3.9% | 14.8% | 10.3% | 12.3% | 1.3% | 65.2% | 3.9% |

Kennedy, Sagae, & Dehghani, 2019; Lin et al., 2018; Warner & Hirschberg, 2012). In Table 1, the overall binary distributions of the three main labels is displayed. Approximately 9% of the entire corpus is either HD or CV, as decided by a majority of the annotators. In addition, 6% of the entire corpus contains vulgar, offensive, or attacking language that does not reference the target's group identity.

In Table 2, the distribution of targeted populations as well as the implicit or explicit framing are displayed. Note that the coverage of these positive percentages are not complete, given that annotations that did not label a document as either HD or CV were not considered, and this proportion was a non-insignificant part of the majority-aggregated positive class. One can see that the two most common target populations were religious/ethnicity (∼29% of HD and ∼14% of CV) and religious (∼22% of HD and ∼19% of CV). The majority of hate-based rhetoric in the GHC are explicit in their rhetoric, though 14% of HD documents conveyed their hatred through subliminal, more context-sensitive ways.

**Classification Analysis**

Hate speech classification is one of the major objectives of the GHC, for a variety of purposes: general purpose detection, for moderating online social platforms; applying

predictive models to other, large datasets for downstream analysis; and post-hoc deconstruction of linguistic features associated with annotated hate speech. In order to establish classification baselines, here we provide cross-validated metrics of performance for three methods which are representative of many of the standard approaches used in NLP: bag of words modeling, dictionary-based measures, and language model fine-tuning. We apply these methods to the HD and VO labels, and leave further analysis of target population and framing to others. We also train models on the "Hate" label, which is the annotator-level union of HD and CV, given that CV labels are too sparse to provide enough signal to train predictive models.

After evaluating each classifier, we apply it to the full dataset of Gab posts released by Gaffney (2018), in order to estimate the distribution of hate-based rhetoric on the entire social network. Using retrained models using the best-fit parameters of the TF–IDF and LIWC models, and the best-performing fine-tuned BERT model weights (by $F_1$), we predicted labels for HD, Hate, and VO over 15,675,294 data points, which were filtered from the original dataset by whether they included at least three valid tokens.

**Analytic Approach**

Our first two methods for representing text operate within the bag of words framework — that is, documents are represented by the (normalized) count of each word in a fixed vocabulary. First, Term Frequency–Inverse Document Frequency (TF–IDF) is a general-purpose method for normalizing word frequencies by their inverse document frequency, which was popularized in information retrieval (Aizawa, 2003) and remains a strong baseline in text prediction (e.g., Joulin, Grave, Bojanowski, & Mikolov, 2017). Second, we use the primary set of psychological dictionaries from the 2015 Linguistic Inquiry and Word Count program (LIWC; Pennebaker, Boyd, Jordan, & Blackburn, 2015), totalling 73 categories of words defined according to psychological or linguistic attributes — e.g., affect, social, part of speech, and topic. LIWC is used for this analysis both for its

established presence — many psychological constructs have been coded and measured through LIWC dictionaries — and the potential intersection between known psychological constructs, as measured through word occurrence, and hate speech. We apply TF–IDF and LIWC normalized word counting to the entire corpus and apply Support Vector Machines (SVM; Cortes & Vapnik, 1995) for classification with these two feature sets. SVMs learn a classification boundary by mapping training data into high-dimensional space, and finding the widest gap between differing classes. Unseen test data is then mapped into the same space, and classified based on their position with regards to the boundary. SVMs are particularly effective statistical learners and are established as a standard in text classification (Joachims, 1998).

In addition to these two standards in text analysis, we apply the emerging state of the art in NLP research, consisting of the transfer of learned linguistic knowledge from language modeling to potentially small-scale downstream applications (Devlin, Chang, Lee, & Toutanova, 2018; J. Howard & Ruder, 2018; Liu, Gardner, Belinkov, Peters, & Smith, 2019; Peters et al., 2018; Radford, Narasimhan, Salimans, & Sutskever, 2018). We apply one of the most successful techniques, fine-tuning "Bidirectional Encoder Representations from Transformers" (BERT; Devlin et al., 2018) using the Pytorch (Paszke et al., 2019) implementation of Transformer language model fine-tuning by "Huggingface"[2] (Wolf et al., 2019). BERT consists of a large neural network (either 12 or 24 layers) which learns an advanced degree of compositionality, syntax, and word meaning through series of predictive language modeling objectives, including predicting missing words in sentences as well as the order of sentences. The efficacy of BERT and similar models is in transfer: model weights are saved, distributed, and subsequently "fine-tuned" on downstream tasks. In fine-tuning, models which have been previously "pre-trained" on massive text corpora, ranging from books, online media, and Wikipedia, are trained on a prediction task such as text classification, refining model weights in order to maximize task performance. In the

[2] `https://github.com/huggingface/transformers`

present work, we adopt this fine-tuning approach, taking BERT models trained on massive text datasets and adjusting them for the classification tasks of this corpus.

### Implementation details

SVM models are fit to each feature set (TF–IDF and LIWC) with the scikit-learn v0.22.1 (Pedregosa et al., 2011) Python 3.6 machine learning library. Default parameters of the "LinearSVC" class (implementing Fan, Chang, Hsieh, Wang, and Lin (2008)'s SVM with linear kernel) were used, except for $C$ (controlling level of L2 regularization) and *class weight* (controlling whether losses are weighted according to class imbalance or not), which were selected based on grid search. BERT models were fine-tuned following the instructions of (BERT; Devlin et al., 2018). First, each post was tokenized as described in the original paper. As around 99% of the posts had less than 100 tokens, thus we kept the first 100 tokens of each post. We used a batch size of 16 and fine-tuned the pre-trained BERT model for 4 epochs. Using one Nvidia 2080 Super GPU, each training epoch took about five minutes. No additional parameter tuning or early stopping was performed, thus further work can determine the maximum predictive performance obtainable through BERT.

## Results

### Classification Performance

All models were evaluated by ten-fold cross-validation. In Table 3, the average and standard deviation of four different metrics — accuracy, precision, recall, and $F_1$ score — are reported for predicting labels of the majority-aggregated dataset. F1 score is the harmonic mean between precision and recall, and is commonly used to evaluate classification on imbalanced data, since accuracy is artificially high in these cases. High precision indicates that a model is more conservative in predicting the positive class, high recall indicates that a model learns a more comprehensive representation of the positive class, and high $F_1$ indicates an ideal balance between the two. Accuracy scores are high

due to the sparsity of the prediction tasks, and are less meaningful than $F_1$.

Table 3
*Mean and standard deviation of $F_1$, precision, recall, and accuracy for predicting HD, VO, and Hate (union of HD and CV) across 10-fold cross validation.*

| Model | Label | $F_1$ | Accuracy | Precision | Recall |
|---|---|---|---|---|---|
| LIWC | HD | 0.26 (0.01) | 0.71 (0.01) | 0.60 (0.03) | 0.17 (0.01) |
| | VO | 0.33 (0.02) | 0.85 (0.01) | 0.58 (0.03) | 0.23 (0.01) |
| | Hate | 0.39 (0.02) | 0.71 (0.01) | 0.56 (0.03) | 0.30 (0.02) |
| TF-IDF | HD | 0.40 (0.02) | 0.84 (0.01) | **0.62 (0.03)** | 0.29 (0.02) |
| | VO | **0.42 (0.03)** | 0.89 (0.01) | **0.65 (0.05)** | 0.31 (0.02) |
| | Hate | 0.53 (0.02) | 0.81 (0.01) | **0.66 (0.02)** | 0.44 (0.01) |
| BERT | HD | **0.44 (0.03)** | **0.92 (0.01)** | 0.46 (0.03) | **0.42 (0.02)** |
| | VO | **0.42 (0.03)** | **0.94 (0.00)** | 0.45 (0.03) | **0.39 (0.03)** |
| | Hate | **0.58 (0.02)** | **0.87 (0.01)** | 0.58 (0.02) | **0.57 (0.02)** |

LIWC and TF–IDF predicted labels are less accurate, though overall more conservative estimates. BERT predicted labels have higher coverage of hate, and are overall more accurate. Within this overall improvement, BERT models exchanged a higher false positive rate (higher precision) for a lower false negative rate (high recall). The low precision of BERT models was not due to an inferior modeling capacity, but merely to the way in which the fine-tuning was performed. SVM models were trained used class weighting, in which the loss from each binary class was weighted inversely according to frequency; i.e., sparse positive classes were prioritized. Performing class weighting while fine-tuning BERT is possible, though non-trivial, and was not performed in this work. It is expected that a class-weighted BERT model would achieve greater F1 performance than the presented model due to an increase in model precision. In the case of VO, the margin between TF–IDF and BERT was slight, suggesting that the presence of offensive and abusive language is sufficient signal for capturing VO. The low performance of LIWC relative to both TF–IDF and BERT can likely be explained by the high dependence of

hate-based rhetoric on non-dictionary words, including social group terms (e.g., "jew"), slurs, and the lower representational power of LIWC in comparison to BERT and other transfer-learning methods. TF–IDF and fine-tuned BERT classifiers were applied to the full Gab corpus, yielding labels for HD, Hate, and VO for 15,675,294 Gab posts. These labels are released as part of the GHC. The binary distributions of each label, produced by each model, are included in Table 4

| | TF–IDF | BERT |
|---|---|---|
| VO | 1.2% | 4.7% |
| HD | 1.6% | 8.0% |
| Hate | 6.7% | 18.3% |

Table 4
*Distribution of labels over the full release of Gab (N = 15,675,294).*

## Data Records

The GHC is available at `https://osf.io/edua3/`. The release of the dataset includes a full record of each annotation ($n$ = 91,967) for all posts, as well as a release of the predicted hate-based rhetoric labels for each classifier reported above. Annotators are anonymized to random, nondescriptive identifiers. Users of the GHC can perform aggregation of the annotations either as reported in the present work (majority vote) or through other techniques. All Gab user information is removed from the released dataset, aside from unique, random identifiers assigned to each user. The release of predicted labels includes document IDs corresponding to the original IDs found in the Pushshift release of the data.

Future additions to the GHC will include annotator information, including demographics, Implicit Association Test (Greenwald, McGhee, & Schwartz, 1998), and measures on attitudes towards hate crime and hate speech censorship and policies (Cabeldue, Cramer, Kehn, Crosby, & Anastasi, 2018).

## Discussion

We introduced the GHC, a large-scale corpus annotated for hate-based rhetoric according to a synthesis of psychological, sociological, computational, and legal definitions of hate speech. At its core, the GHC represents the operationalization of hate speech as a socio-psychological construct, defined by the functional role of hate speech within existing social hierarchies. It complements and extends previous annotation data projects, notably Davidson et al. (2017), Waseem and Hovy (2016), and de Gibert et al. (2018), through the comprehensiveness of its typology, its theoretical basis, and its novel data domain. The size of the corpus makes it particularly suitable for computational analyses, whether their goals are detection or analysis, given the high sample complexity of state of the art models in NLP. Most importantly, the GHC is publicly available, creating the possibility of collaboration between computer scientists, psychologists, sociologists, political scientists, legal scholars, public policy experts, and others who aim to measure, understand, and/or mitigate hate speech and prejudice in online media.

The data annotation paradigm which we present here represents an emerging approach in computational social science for studying compositional textual phenomena — i.e., socio-psychological constructs in language that are not captured by the presence or non-presence of a particular keyword, or set of keywords. In computational linguistics, annotation is an established approach for studying syntactic and semantic structure in natural language (e.g., Banarescu et al., 2013; Hovy, Marcus, Palmer, Ramshaw, & Weischedel, 2006; Marcus, Santorini, & Marcinkiewicz, 1993). However, annotation can also be a "psychological" task. Annotation is used in sentiment analysis and opinion mining (Pang, Lee, et al., 2008; Pang, Lee, & Vaithyanathan, 2002), in which subjective aspects of text are manually labeled and used for training classifiers (e.g., Agarwal, Xie, Vovsha, Rambow, & Passonneau, 2011). Similarly, here we performed psychologically-informed expert annotation of hate-based rhetoric, an inherently subjective phenomenon. Large-scale text annotation is a relatively new paradigm in psychology (see

Hoover et al., 2020). It bears some similarities to established annotation paradigms in computational linguistics — notably, the evaluation of inter-annotator agreement and the generation of training labels for machine learning algorithms — but it expands on established methods in important ways, notably in the amount of subjectivity involved in producing annotations and scientific applications of the corpus beyond classification.

We emphasize these two distinctions of annotation of socio-psychological phenomena — subjectivity and scientific applications — for future work dealing with the GHC. Annotating socio-psychological phenomena is not the same as measuring objective linguistic phenomena (e.g., tagging the parts of speech of words in a sentence), in which disagreement can be confidently attributed to random variation in the data or poorly specified instructions. In considering the variation in individuals' treatment of different social groups' warmth and competence (Fiske, Cuddy, & Glick, 2007), Mostafazadeh Davani et al. (2020) provided initial evidence that models trained on the GHC take on human-like social stereotypes, evaluating hate speech differently depending on the stereotypes associated with the named target group. Future research can proceed in this vein, both quantifying and correcting variability in annotations that are driven by differences in psychological characteristics.

Remaining questions about the GHC are largely technical, and have to do with building, evaluating, and deconstructing models of hate-based rhetoric. Hate speech models are known to suffer from a "bias" problem (Wiegand et al., 2019), i.e., they contain the imbalances and irregularities that are seen in the language distribution of hate speech datasets. The same is true for the GHC: Kennedy, Jin, Mostafazadeh Davani, Dehghani, and Ren (2020) found that state of the art models trained on the GHC disproportionately associate the presence of group-level terms with hate-based rhetoric. Terms like "jew" and "black", when found in an innocuous (e.g., informative, conversational, etc.) context, influenced models to judge a piece of text as hate speech. Correcting this oversensitivity to group terms was done by applying term-level regularization to group terms through

post-hoc explanation algorithms (Jin, Du, Wei, Xue, & Ren, 2019); in other words, models were constrained to only *partially* use group terms in their representations. The result was that models were better able to capture hate-based rhetoric and less likely to misjudge innocuous occurrences of group terms. In general, further research into identifying and mitigating bias in hate speech models is supported by the GHC.

The desired result unbiased, high-powered classifiers trained on the GHC is that they can be confidently applied to algorithmically label data that are not yet annotated. Transferring annotated information in this way can be used to infer the distribution of hate-based rhetoric in various domains, such as mainstream social media, transcripts of political speeches or other political documents, news articles and reporting, and the rest of the Gab domain. For example, Hoover et al. (2019) trained hate speech classifiers on a subset of the GHC and applied them to the entire Gab corpus, thereby analyzing the correlation of moral and hateful sentiments across an exhaustive sample of hate. Previous work in computational social science which has utilized this approach sought to understand the effect of moral rhetoric in online social movements, particular as they relate to violence at protests (Mooijman, Hoover, Lin, Ji, & Dehghani, 2018).

More recently, Atari et al. (2020) used predicted labels using the GHC to determine the relationship between group homophily in terms of moral values and the intensity of hate-based rhetoric in Gab. These authors also annotated a small dataset from misogynist subreddit (Incels), finding that GHC-trained classifiers can accurately detect hate speech in other platforms such as Reddit. By annotating large numbers of posts from these platforms and training neural networks showcased in this paper, they automatically labeled millions of Gab and Incels posts.

## Conclusion

Hate speech, particularly its online presence, is a harmful and endangering social act that clearly requires coordinated research across disciplines and sectors. The present

work contributes to this coordination by developing a large-scale resource for computational modeling and analysis. The GHC will benefit computational modeling research, facilitating a more theoretically-informed quantification of hate speech, providing a large-scale resource with high quality annotation according to a multi-level typology, and clearly identifying modeling challenges. The questions that are as yet unanswered regarding the subjectivity of annotation can be addressed by psychological research, where existing theoretical frameworks can be applied in order to attribute systematic differences in annotation. The present work supports the emerging possibility of studying the psychology of prejudice and hate speech through language. These are practically difficult phenomena to study empirically, and language — including unobtrusive observations of anonymous online posts — offers a cheap and accessible window to reaching radicalized individuals as well as their rhetorical tools.

References

Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. J. (2011). Sentiment analysis of twitter data. In *Proceedings of the workshop on language in social media (lsm 2011)* (pp. 30–38).

Aizawa, A. (2003). An information-theoretic perspective of tf–idf measures. *Information Processing & Management*, *39*(1), 45–65.

Allport, G. W., Clark, K., & Pettigrew, T. (1954). The nature of prejudice.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders (dsm-5®)*. American Psychiatric Pub.

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

Atari, M., Mostafazadeh Davani, A., Kogon, D., Kennedy, B., Saxena, N. A., Anderson, I., & Dehghani, M. (2020). *Extremists of a feather, hate together: Morally homogeneous networks and use of hateful rhetoric.*

Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., ... Schneider, N. (2013). Abstract meaning representation for sembanking. In *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse* (pp. 178–186).

Bell, H., Kulkarni, S., & Dalton, L. (2003). Organizational prevention of vicarious trauma. *Families in society*, *84*(4), 463–470.

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

Boeckmann, R. J., & Liew, J. (2002). Hate speech: Asian american students' justice judgments and psychological responses. *Journal of Social Issues*, *58*(2), 363–381.

Bradbury, S. (2018). Timeline of terror: A moment-by-moment account of squirrel hill mass shooting. *Pittsburgh Post-Gazette*. Retrieved 2018-10-28, from

https://www.post-gazette.com/news/crime-courts/2018/10/28/TIMELINE-20-minutes-of-terror-gripped-Squirrel-Hill-during-Saturday-synagogue-attack/stories/201810280197

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

Byrt, T., Bishop, J., & Carlin, J. B. (1993). Bias, prevalence and kappa. *Journal of clinical epidemiology*, *46*(5), 423–429.

Cabeldue, M. K., Cramer, R. J., Kehn, A., Crosby, J. W., & Anastasi, J. S. (2018). Measuring attitudes about hate: Development of the hate crime beliefs scale. *Journal of interpersonal violence*, *33*(23), 3656–3685.

*Chaplinsky v. new hampshire* (Vol. 315) (No. No. 255). (1942). Supreme Court.

*Community standards: 12 hate speech.* (2020). https://www.facebook.com/communitystandards/hate_speech. (Accessed: 2020-02-07)

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273–297.

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

de Gibert, O., Perez, N., García-Pablos, A., & Cuadros, M. (2018). Hate speech dataset from a white supremacy forum. *arXiv preprint arXiv:1809.04444*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dixon, L., Li, J., Sorensen, J., Thain, N., & Vasserman, L. (2018). Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 aaai/acm conference on ai, ethics, and society* (pp. 67–73).

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., & Lin, C.-J. (2008). Liblinear: A

library for large linear classification. *Journal of machine learning research*, *9*(Aug), 1871–1874.

Fiske, S. T., Cuddy, A. J., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in cognitive sciences*, *11*(2), 77–83.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, *76*(5), 378.

Gaffney, G. (2018). *Pushshift gab corpus.* `https://files.pushshift.io/gab/`. (Accessed: 2019-5-23)

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

Garten, J., Kennedy, B., Sagae, K., & Dehghani, M. (2019). Measuring the importance of context when modeling language comprehension. *Behavior research methods*, *51*(2), 480–492.

*German criminal code.* (1998). https://www.gesetze-im-internet.de/englisch_stgb/englisch_stgb.html.

Glaser, J., Dixit, J., & Green, D. P. (2002). Studying hate crime with the internet: What makes racists advocate racial violence? *Journal of Social Issues*, *58*(1), 177–193.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, *74*(6), 1464.

Grey Ellis, E. (2016). On gab, an extremist-friendly site, pittsburgh shooting suspect aired his hatred in full. *WIRED*. Retrieved 2016-09-14, from `https://www.wired.com/2016/09/gab-alt-rights-twitter-ultimate-filter-bubble/`

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

*Hateful conduct policy.* (2020). `https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy`.

(Accessed: 2020-02-07)

*Hate speech policy.* (2020). https://support.google.com/youtube/answer/2801939. (Accessed: 2020-02-07)

Herek, G. M., Gillis, J. R., & Cogan, J. C. (1999). Psychological sequelae of hate-crime victimization among lesbian, gay, and bisexual adults. *Journal of consulting and clinical psychology*, *67*(6), 945.

Hern, A. (2019). Ex-facebook worker claims disturbing content led to ptsd. *The Guardian*. Retrieved 2019-12-04, from https://www.theguardian.com/technology/2019/dec/04/ex-facebook-worker-claims-disturbing-content-led-to-ptsd

Hoover, J., Atari, M., Davani, A. M., Kennedy, B., Portillo-Wightman, G., Yeh, L., ... Dehghani, M. (2019). Bound in hatred: The role of group-based morality in acts of hate.

Hoover, J., Portillo-Wightman, G., Yeh, L., Havaldar, S., Davani, A. M., Lin, Y., ... others (2020). Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment. *Social Psychological and Personality Science*.

Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., & Weischedel, R. (2006). Ontonotes: the 90% solution. In *Proceedings of the human language technology conference of the naacl, companion volume: Short papers* (pp. 57–60).

Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

Howard, J. W. (2019). Free speech and hate speech. *Annual Review of Political Science*, *22*, 93–109.

Jin, X., Du, J., Wei, Z., Xue, X., & Ren, X. (2019). Towards hierarchical importance attribution: Explaining compositional semantics for neural sequence models. *arXiv preprint arXiv:1911.06194*.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In *European conference on machine learning* (pp. 137–142).

Joulin, A., Grave, É., Bojanowski, P., & Mikolov, T. (2017). Bag of tricks for efficient text classification. In *Proceedings of the 15th conference of the european chapter of the association for computational linguistics: Volume 2, short papers* (pp. 427–431).

Kaplan, J. B., Bergman, A. L., Christopher, M., Bowen, S., & Hunsinger, M. (2017). Role of resilience in mindfulness training for first responders. *Mindfulness*, *8*(5), 1373–1380.

Kennedy, B., Jin, X., Mostafazadeh Davani, A., Dehghani, M., & Ren, X. (2020). *Contextualizing hate speech classifiers with post-hoc explanation.*

Kleim, B., & Westphal, M. (2011). Mental health in first responders: A review and recommendation for prevention and intervention strategies. *Traumatology*, *17*(4), 17–24.

Leets, L. (2002). Experiencing hate speech: Perceptions and responses to anti-semitism and antigay speech. *Journal of social issues*, *58*(2), 341–361.

Levin, S. (2017). Moderators who had to view child abuse content sue microsoft, claiming ptsd. *The Guardian*. Retrieved 2017-01-11, from `https://www.theguardian.com/technology/2017/jan/11/microsoft-employees-child-abuse-lawsuit-ptsd`

Lin, Y., Hoover, J., Portillo-Wightman, G., Park, C., Dehghani, M., & Ji, H. (2018). Acquiring background knowledge to improve moral value prediction. In *2018 ieee/acm international conference on advances in social networks analysis and mining (asonam)* (pp. 552–559).

Liu, N. F., Gardner, M., Belinkov, Y., Peters, M., & Smith, N. A. (2019). Linguistic knowledge and transferability of contextual representations. *arXiv preprint arXiv:1903.08855*.

Ludick, M., & Figley, C. R. (2017). Toward a mechanism for secondary trauma induction and reduction: Reimagining a theory of secondary traumatic stress. *Traumatology*, *23*(1), 112.

Marcus, M., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a large annotated

corpus of english: The penn treebank.

Matsakis, L. (2018). Pittsburgh synagogue shooting suspect's gab posts are part of a pattern. *WIRED*. Retrieved 2018-10-27, from `https://www.wired.com/story/pittsburgh-synagogue-shooting-gab-tree-of-life/`

Matsuda, M. J. (1989). Public response to racist speech: Considering the victim's story. *Michigan Law Review*, *87*(8), 2320–2381.

May, C. L., & Wisco, B. E. (2016). Defining trauma: How level of exposure and proximity affect risk for posttraumatic stress disorder. *Psychological trauma: theory, research, practice, and policy*, *8*(2), 233.

McDevitt, J., Balboni, J., Garcia, L., & Gu, J. (2001). Consequences for victims: A comparison of bias-and non-bias-motivated assaults. *American behavioral scientist*, *45*(4), 697–713.

Mondal, M., Silva, L. A., & Benevenuto, F. (2017). A measurement study of hate speech in social media. In *Proceedings of the 28th acm conference on hypertext and social media* (pp. 85–94).

Mooijman, M., Hoover, J., Lin, Y., Ji, H., & Dehghani, M. (2018). Moralization in social networks and the emergence of violence during protests. *Nature Human Behaviour*, *2*, 389–396.

Mostafazadeh Davani, A., Atari, M., Kennedy, B., Havaldar, S., & Dehghani, M. (2020). *Hatred is in the eye of the annotator: Hate speech classifiers learn human-like social stereotypes.*

Müller, K., & Schwarz, C. (2019). Fanning the flames of hate: Social media and hate crime. *Available at SSRN 3082972*.

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., & Chang, Y. (2016). Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web* (pp. 145–153).

Olteanu, A., Castillo, C., Boy, J., & Varshney, K. R. (2018). The effect of extremist

violence on hateful speech online. *arXiv preprint arXiv:1804.05704*.

Pang, B., Lee, L., et al. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, *2*(1–2), 1–135.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the acl-02 conference on empirical methods in natural language processing-volume 10* (pp. 79–86).

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... others (2019). Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems* (pp. 8024–8035).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of liwc2015* (Tech. Rep.).

Perez, L. M., Jones, J., Englert, D. R., & Sachau, D. (2010). Secondary traumatic stress and burnout among law enforcement investigators exposed to disturbing media images. *Journal of Police and Criminal Psychology*, *25*(2), 113–124.

Perry, B. (2002). *In the name of hate: Understanding hate crimes*. Routledge.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf*.

*Rav v. st. paul* (Vol. 505) (No. No. 90-7675). (1992). Supreme Court.

Roose, K. (2018). On gab, an extremist-friendly site, pittsburgh shooting suspect aired his hatred in full. *The New York Times*. Retrieved 2018-10-28, from

https://www.nytimes.com/2018/10/28/us/gab-robert-bowers-pittsburgh-synagogue-shootings.html

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

Rozin, P. (2001). Social psychology and science: Some lessons from solomon asch. *Personality and Social Psychology Review*, *5*(1), 2–14.

Schmidt, A., & Wiegand, M. (2017). A survey on hate speech detection using natural language processing. In *Proceedings of the fifth international workshop on natural language processing for social media* (pp. 1–10).

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

Wagaman, M. A., Geiger, J. M., Shockley, C., & Segal, E. A. (2015). The role of empathy in burnout, compassion satisfaction, and secondary traumatic stress among social workers. *Social work*, *60*(3), 201–209.

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

Warner, W., & Hirschberg, J. (2012). Detecting hate speech on the world wide web. In *Proceedings of the second workshop on language in social media* (pp. 19–26).

Waseem, Z., Davidson, T., Warmsley, D., & Weber, I. (2017). Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the first workshop on abusive language online* (pp. 78–84).

Waseem, Z., & Hovy, D. (2016). Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the naacl student research workshop* (pp. 88–93).

Wiegand, M., Ruppenhofer, J., & Kleinbauer, T. (2019). Detection of abusive language: the problem of biased datasets. In *Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: Human language*

*technologies, volume 1 (long and short papers)* (pp. 602–608).

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... others (2019). Transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Wulczyn, E., Thain, N., & Dixon, L. (2017). Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th international conference on world wide web* (pp. 1391–1399).

Ybarra, M. L., Mitchell, K. J., Wolak, J., & Finkelhor, D. (2006). Examining characteristics and associated distress related to internet harassment: findings from the second youth internet safety survey. *Pediatrics*, *118*(4), e1169–e1177.

# Appendix

## Coding Manual

The below text was originally written for the purpose of defining hate-based rhetoric and training annotators. The original version is available at `https://psyarxiv.com/hqjxn/`.

### Defining Hate-Based Rhetoric

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.

## 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[3].

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:

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

1. “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."

2. 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 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: Explicit or Implicit.** 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.

**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"), or both. If none apply, the document is to be considered NH ("Not Hateful") 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.

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.

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), taken from Davidson et al. (2017):

> 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

Group tags are used to identify which types of groups are targeted by the hateful rhetoric. Select all that 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.

A document 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).

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.

**Training 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 contain hateful rhetoric and abusive language**.

Table A1

*Examples illustrating hate-based rhetoric typology and coding as agreed by authors and annotators*

| **Text** | **Labels** |
|---|---|
| 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 |
| #BLMwants you to think black crime stats only exist because blacks are targeted more often.Easy. Just walk into literally any gas station and ask which race makes them worry.Black targeting is an effect, not a cause, which destroys the entire #Leftistnarrative | HD RAE IM |
| #MiddleEasternanimal kills 8, injures 15 in a city that won't learn its lesson about #Islam.This creature should be lynched, dragged til dead, and his head on a pike on the Brooklyn Bridge. | HD CV REL NAT EX |
| The problem is that we're not speaking a language the #IslamoFascist-sunderstand.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 |
| The ZioJesuit cultural marxist anti-white Christian killers now working within our white governments globally should be shot for espionage and treason with their non-white invaders expelled from within our nations immediately for the sake of our future white Christian posterity. Anything else is cowardice, self-eradication & continued genocide. | CV RAE EX |
| Ok so apparently they will be sent back to shitholes.Like I care? Nope. | HD, NAT, IM |

Table A2
*Examples (continued from Table A1)*

| **Text** | **Labels** |
|---|---|
| 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. Facts are Facts. Let's just stop being stupidly ignorant and not wait the complete extinction of our brothers in SA to understand it ! #WhiteGenocide #SaveTheBoer | HD RAE NAT EX |
| 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 |