PREPRINT – ARTICLE IN PRESS: Kennedy, B., Atari, M., Mostafazadeh Davani, A., Yeh, L., Omrani, A., Kim, Y., Coombs, K., Havaldar, S., Portillo-Wightman, G., Gonzalez, E., Hoover, J., Azatian, A., Hussain, A., Lara, A., Cardenas, G., Omary, A., Park, C., Wang, X., Wijaya, C., Zhang, Y., Meyerowitz, B., Dehghani, M. Introducing the Gab Hate Corpus: Defining and applying hate-based rhetoric to social media posts at scale. *Lang Resources & Evaluation (2022).* Available at `https://doi.org/10.1007/s10579-021-09569-x`.

Introducing the Gab Hate Corpus: Defining and applying hate-based rhetoric to social media posts at scale

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

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

[2]Department of Psychology, University of Southern California

[3]Gould School of Law and Department of Political Science & International Relations, University of Southern California

Author Note

Correspondence regarding this article should be addressed to Brendan Kennedy (btkenned@usc.edu) or Morteza Dehghani (mdehghan@usc.edu) 362 S. McClintock Ave, Los Angeles, CA 90089-161. All data have been publicly deposited to the Open Science Framework repository (`https://osf.io/edua3/`). This research was sponsored by NSF CAREER BCS-1846531 to MD.

Abstract

We present the Gab Hate Corpus (GHC), consisting of 27,665 posts from the social network service `gab.com`, each annotated for the presence of "hate-based rhetoric" by a minimum of three annotators. Posts were labeled according to a coding typology derived from a synthesis of hate speech definitions across legal precedent, previous hate speech coding typologies, and definitions from psychology and sociology, comprising hierarchical labels indicating dehumanizing and violent speech as well as indicators of targeted groups and rhetorical framing. We provide inter-annotator agreement statistics and perform a classification analysis in order to validate the corpus and establish performance baselines. The GHC complements existing hate speech datasets in its theoretical grounding and by providing a large, representative sample of richly annotated social media posts.

Keywords: Hate Speech; Text Classification; Social Media; Social Science

Introducing the Gab Hate Corpus: Defining and applying hate-based rhetoric to social media posts at scale

Introduction

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 additional attention given the attacker's vendetta against the Jewish people, which he documented on his account on the social media site "Gab" (Roose, 2018). Gab, 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). As a social media platform and open space for the sharing of dangerous ideas, it has helped extremists organize and carry out an attempted insurrection in the U.S. capitol (Romero, 2021). Gab is one of several online communities producing disproportionate amounts of hateful and abusive rhetoric (Matsakis, 2018).

As hateful groups and behaviors grow in prominence in online spaces, there is an acute need for more effective combative strategies. This not only includes automatic detection and removal of harmful posts — commonly addressed in the Natural Language Processing (NLP) community — but also broad measures such as "counterspeech" (Gagliardone, Gal, Alves, and Martinez, 2015) and scientific research into understanding the psychological antecedents of prejudice and the motivations leading to the production of harmful content.

However, these strategies are often pursued in isolation. Legal scholarship focuses on the correct definition of hate speech and the proper course of action towards combatting it, with little agreement among countries. Previous research in sociology and political psychology has focused on explaining the formation of prejudice and hate groups, but struggle with limited access to naturalistic instances of hate speech and hate crime. And most NLP studies of hate speech — notably the construction of labeled datasets that operationalize hate speech — are unconnected to the rich literature on hate speech that predates the recent increase of hatred in online spaces.

In general, a more integrated approach to hate speech can result in improved understanding of hate groups and hateful behaviors, detection of hate speech, and development of policy based on data-driven insight. In particular, the present work focuses on the problem of hate speech operationalization in NLP. Whereas previous works often use ad hoc coding rules, generic definitions, or non-expert annotation (for a discussion of the importance of expertise in annotating hate speech, see Waseem and Hovy, 2016), we propose an operationalization of hate speech that is inclusive with respect to social scientific and legal precedent, which we call *hate-based rhetoric*. Our operationalization of hate speech, which is our first main contribution, is grounded in legal precedent surrounding hate speech and censorship beyond just the United States; for example, the German Criminal code includes "incitement to hatred," which is more inclusive and sensitive to cultural context than similar laws in the United States. In addition, our operationalization and associated coding typology is influenced by previous scholarship in sociology and psychology on understanding the complex motivations and function of hate speech, in particular hateful ideologies and groups.

Our second main contribution is to provide expert annotations of hate-based rhetoric, based on a dedicated coding typology, for a large sample of Gab posts. This annotation process, involving considerable human effort and coordination, was accompanied by efforts to monitor annotators for harms suffered during annotation of hateful content. Finally, our third contribution is to report results from a thorough set of classification experiments, establishing performance baselines for our top-level categories of Human Degradation ($F_1 = 0.49$), Calls for Violence ($F_1 = 0.32$), and Vulgar/Offensive speech ($F_1 = 51.4$) by fine-tuning pretrained Transformer models.

In its typological richness, sampling representativeness, and scale, the "Gab Hate Corpus" (GHC) provides a valuable resource for the development and validation of hate speech classifiers. Just as important as providing a practical tool for NLP research, the GHC has the potential for increasing collaboration among the social sciences, legal scholars and policy makers, and computational researchers, due to its heightened theoretical grounding and its potential use to study hate groups in online spaces (e.g., white supremacists on Gab).

The remainder of this article is organized as follows. First, we review relevant literature on the aforementioned study of hate speech in the legal and social science communities, computational strategies for classifying hate speech, and the particular task of identifying data domains for the study of hate speech. Then, consulting legal precedent on hate speech, in particular precedent outside of the U.S., a new operationalization based on notions of "incitement to hatred" is presented. Coding instructions were designed according to these legal sources as well as sociological perspectives on the motivations and functional purposes of hatred and prejudice. In Section "Methodology," this operationalization was applied to a large sample of Gab posts (N = 27,655). We document the training of a cohort of undergraduate research assistants, the inter-annotator reliability measures, and the resulting label distributions. For the benefit of future work dedicated to similar annotation of hate speech, we pay special attention to the issues involved in exposing annotators to harmful content. Lastly, our Section on "Experiments" reports the results of a set of baseline classification experiments. This includes classification performance for each major label using bag of words features, language model fine-tuning, an analysis of feature importance for these models, and misclassification analysis.

Background

Hate Speech Scholarship

Hate speech has historically been an ambiguous term, with many interpretations in casual as well as legal usage (Sellars, 2016). It is frequently at the center of the "free speech" discourse, specifically concerning whether citizens have a right to free expression, and thus hate speech ought to be protected, or have certain categories of harmful speech off limits (Howard, 2019). Critically, countries differ in their treatment of hate speech, with the U.S. protecting more categories of speech than other countries, such as Germany, Australia, and the Netherlands. In a later section, we review these differing legal treatments of hate speech across countries in greater detail.

Towards understanding and explaining hate speech, the study of hatred and prejudice in sociology and psychology is both established and actively evolving. Research on the psychology of prejudice extends at least as far as Allport's 1954 treatise on the subject. In sociology research prejudicial movements and ideologies (e.g., anti-black racism in the U.S. South) are analyzed in terms of origins, goals, appeal, and influence (Perry, 2002). In political psychology, one prominent study was Glaser, Dixit, and Green's attempt to interactively analyze the factors that drive the formation of hate groups (2002). Specifically, Glaser et al. studied the attitudes of white supremacists via an anonymous online forum, finding that prejudice is driven by cultural factors (e.g., an aversion to interracial marriage) versus economic factors (competition over resources). A common theme of these, and similar, works is the focus on the motivations of hate groups, rather than the identification of instances of hate or hate speech.

Current research on hate speech classification has to do with the relatively new phenomenon of *online* hate speech. The movement of hate groups to the Internet has spurred a mobilization of novel research and the design of new countermeasures. Of note, the United Nations Educational, Scientific and Cultural Organization (UNESCO; Gagliardone et al., 2015) delivered a program for countering online hate speech, outlining its particular challenges. For example, they observe that online hate speech does not occur necessarily in a single country, thus jurisdiction over online offences is less than clear. Technology companies including Facebook (Facebook, 2020), Twitter (Twitter, 2020), and Google (Google, 2020) have posted official policies on hate speech and abusive language that sometimes go beyond legal restrictions.

The rise of online hate speech has also led to a new wave of data-driven research into hatred, prejudice, and online radicalism. Previously, such scholarship was rendered impossible by a lack of access to relevant data: hate-related behaviors are, fortunately, rare in a statistical sense, and cannot be studied in laboratory settings as individuals are likely to self-censor their true attitudes towards participation in hateful activities (e.g., Kuklinski, Cobb, and Gilens, 1997). Recent data-driven scientific works that study social media data include Müller and Schwarz (2019), which studied the ways in which prejudice manifests in violence in a temporal analysis of anti-refugee activity on Facebook; Atari, Davani, Kogon, Kennedy, Saxena, Anderson, and Dehghani (2021), which included an analysis of the relationship between moral homogeneity in an online social network and the rate of posting hate speech; and Mathew, Illendula, Saha, Sarkar, Goyal, and Mukherjee (2020) traced the temporal and network structure of hate speech in an online social network, Gab (see below).

Hate Speech Classification

Over the past decade, hate speech[1] classification has risen in prominence among NLP tasks (Schmidt and Wiegand, 2017), with considerable subtask diversity (Waseem, Davidson, Warmsley, and Weber, 2017), organization of events and special issue workshops (Akiwowo, Vidgen, Prabhakaran, and Waseem, 2020), and a wide variety of datasets for classifier training and evaluation (Poletto, Basile, Sanguinetti, Bosco, and Patti, 2020). In addition to providing data for training and evaluating classifiers, these datasets have begun to be used to monitor models' tendencies to capture biases against protected groups from training data (Dixon, Li, Sorensen, Thain, and Vasserman, 2018; Mostafazadeh Davani, Atari, Kennedy, Havaldar, and Dehghani, 2020).

The labels in these datasets, whether from manual annotation or collection of metadata, represent varying operationalizations of hate speech. For example, Warner and Hirschberg (2012) labeled social media posts for hate speech by identifying "harmful

[1] We will use "hate speech" to refer to the broad set of tasks in this category, including abusive and toxic language.

stereotypes" against minorities and protected groups; Waseem and Hovy (2016) were partially motivated by theories of white privilege in composing a decision list for identifying hate speech, focusing on speech which "undercut and silenced [minorities]" (p. 89); and Nobata, Tetreault, Thomas, Mehdad, and Chang (2016) operationalized abusive language as content flagged by official moderators. Whether motivated by a notion of harmful stereotypes, a context-sensitive reading of minority oppression in language, or metadata provided by content moderators, classification analysis is directly dependent on the chosen operationalization of hate speech.

The relevance and utility of hate speech datasets is not only determined by the quality of the operationalization, but also by the domain of language contained in a given dataset. Given that our work is partially aimed at connecting data-driven social scientific research on hate to NLP studies of hate speech classification, we focus on domains that are particularly salient to the former. Specifically, rather than annotating data from general social media platforms (e.g., Twitter), we focus on domains that have disproportionate amounts of hate speech, hateful ideology, and hate communities.

The domain we chose, Gab, is not strictly devoted to hate speech, though its censorship policies (of which there are few; see GAB AI., INC's Terms of Service, 2020) and emphasis on "free speech" have contributed to the growth of Gab as a haven for political extremists, the "alt-right," conspiracy theorists, and violent actors (Andrews, 2021). As such, annotating posts from Gab can help to understand and mitigate the risks of violence and extremism posed by the hate-based groups which use Gab.

Recent work in hate speech detection has also begun to focus on Gab. Qian, Bethke, Liu, Belding, and Wang (2019) studies the particular problem of learning when to intervene in hate speech, using data from Gab and Reddit. The authors used the commenting structure of both platforms in order to study the effectiveness of intervening comments (e.g., "This language is offensive and unnecessary") in response to hate speech, providing a crowd-sourced dataset of hate speech labels and generated interventions. Mathew, Dutt, Goyal, and Mukherjee (2019) considered how hate speech spreads across a network, analyzing millions of automatically tagged messages across a large network of Gab users. In the present work, we present a richly annotated dataset of Gab posts that can complement and extend these previous works on the Gab domain, which serve to improve hate speech detection in general as well as analyze the community using Gab to spread hatred.

Hate-Based Rhetoric

Legal precedents surrounding hate speech in Germany, Australia, the Netherlands, and other countries that criminalize speech beyond the standards in the United States, which is well-known to protect many categories of speech (Howard, 2019), are the basis for our operationalization of hate speech. For example, Holocaust denial is criminalized in Germany given its “incitement to hatred.” By relying on the more restrictive laws outside the United States, we cast a wider net than relying on laws and cultural conceptions that place a higher emphasis on the freedom of speech. As a result, our operationalization captures speech that uses cultural context, harmful stereotypes, and background knowledge.

In the development of a coding typology and training guide for annotation, we relied on the language used in sociology research (e.g., the dehumanizing function of prejudice) and the typological structure in previous annotation studies. Our hierarchical typology with high-level categories (indicating the presence of hate-based rhetoric), indicators of targeted populations (e.g., race or ethnicity), differentiation between hate-based rhetoric and merely vulgar or aggressive language, and differentiation between implicit and explicit rhetoric.

Operationalizing Legal and Sociological 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

(Howard, 2019; Sellars, 2016). Indeed, the term “hate speech” was coined fairly recently, with Matsuda (1989) arguing 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). The term today is more of a cultural term that is implicitly understood, despite its legal orientation.

The difficulties in settling on one particular definition of hate speech mean that annotating without formalization (i.e., instructing annotators to label texts as “hate speech” without further detail or explanation of how hate speech is defined) is liable to produce different interpretations of what hate speech actually means. Compounding this issue, notions of hate speech in the United States, which is often the studied population (i.e., U.S. Twitter users) or the country of origin for researchers, are markedly different from other countries. In the United States, which prides itself on its freedoms, particularly those of speech (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 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 shared by 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), as countering such speech requires first recognizing the motivating factors leading to its production in the first place.

Like Germany and other countries, 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 are culturally embedded. 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.

In summary, hate speech is fundamentally embedded within the existing cultural and social context in which it occurs. This is further examined in sociology, in which hate crime has been studied in terms of the relationships between individual perpetrators and larger social movements: “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). Further, hate speech is defined by perpetrators’ desire 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. With this prior scholarship in mind, and given the prior work in hate speech detection research, we define hate-based rhetoric as follows:

> *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.*

Typology and Coding Procedure

One of the main contributions of this paper is a hierarchical coding typology, which was developed to facilitate a more consistent, informed annotation across annotators as well as produce structured information, such as the systematic categorization of target populations (also see Mondal, Silva, and Benevenuto, 2017) and framing effects (also see Olteanu, Castillo, Boy, and Varshney, 2018; Waseem et al., 2017). The full version of the typology used to train annotators, which includes our definitions, explanations of categories, and examples, has previously been made public via online preprint,[2] and is provided in its entirety as part of the Supplemental Materials. The essential components of the typology and the associated process followed by annotators are shown in Figure 1.

Top-level categories. 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. A necessary requirement under this definition is that hate-based rhetoric is explicitly targeting a social group (or an individual by virtue of their social group). Two general categories of such speech were assessed in the present corpus: first, “assaults on human dignity” (HD) broadly include 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). “Calls for violence” (CV) are an explicit call for, or endorsement of, violence on the basis of these descriptions or justifications.

The separation of HD and CV was done according to legal precedent, best

[2] Available as version 1 of the preprint at https://psyarxiv.com/hqjxn/

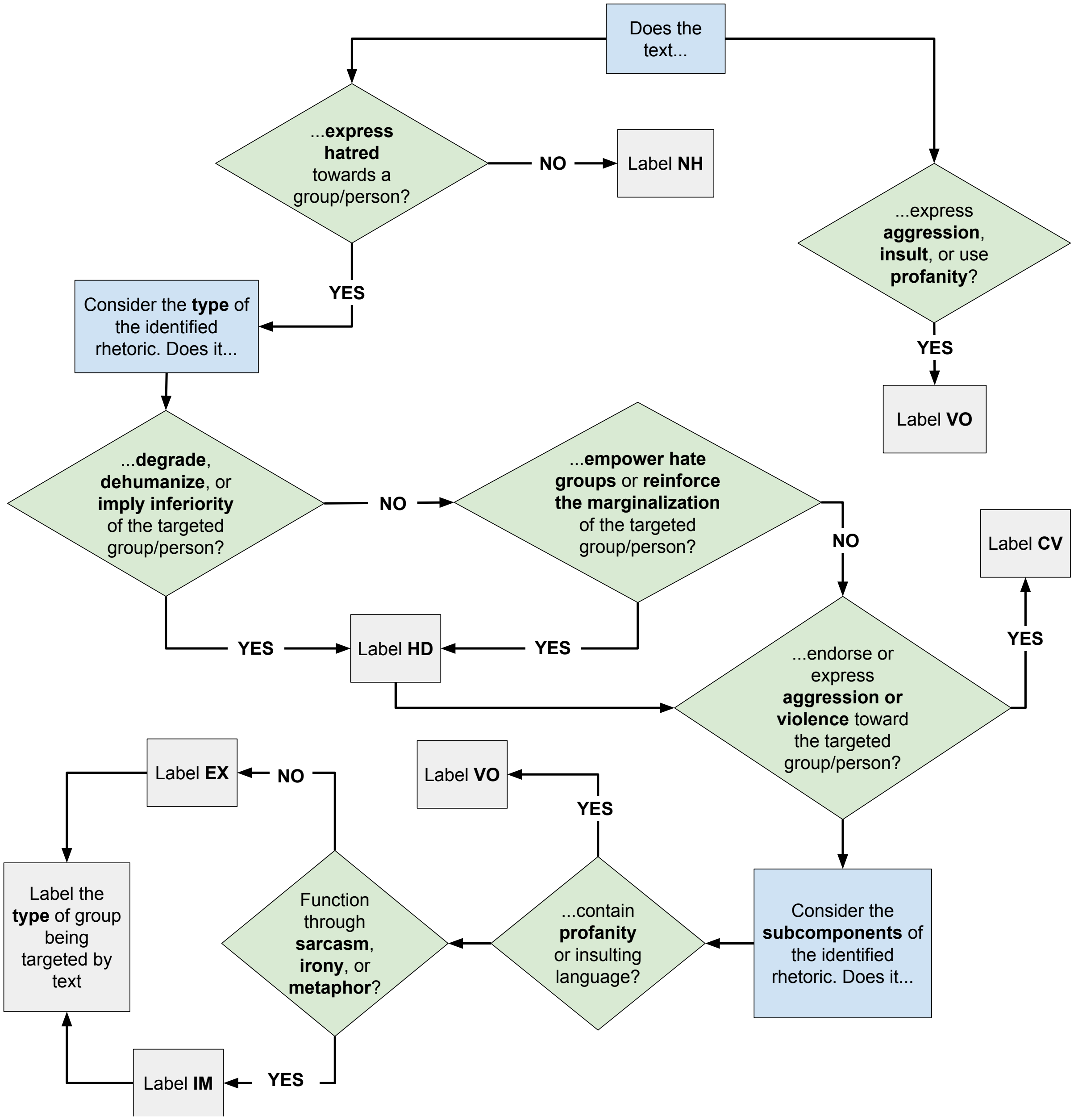
Does the text...
...express hatred towards a group/person?
NO
Label NH
YES
Consider the type of the identified rhetoric. Does it...
...degrade, dehumanize, or imply inferiority of the targeted group/person?
NO
...empower hate groups or reinforce the marginalization of the targeted group/person?
YES
Label HD
YES
NO
...express aggression, insult, or use profanity?
YES
Label VO
Label CV
YES
...endorse or express aggression or violence toward the targeted group/person?
Consider the subcomponents of the identified rhetoric. Does it...
...contain profanity or insulting language?
YES
Label VO
Function through sarcasm, irony, or metaphor?
NO
Label EX
YES
Label IM
Label the type of group being targeted by text

Figure 1. Workflow of annotating a document for hate-based rhetoric.

summarized by the following two-class specification given by 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.

Targeted population for HD and CV. 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.

Vulgarity and/or offensive language. 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). Notably, this last category is often not covered under the U.S. laws or international law.

Framing HD and CV as implicit vs. explicit. 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.

Relation to Existing Datasets

Hate-based rhetoric shares similarities with several previous operationalizations used to produce hate speech datasets, though with several important distinctions. Previous works that operationalize hate speech most similarly to hate-based rhetoric are listed in Table 1. In all cases, we argue that operationalizations are underspecified, relying on high-level definitions or short descriptions of target labels. Many datasets are not included in this comparison, such as those that target different phenomena (e.g., abusive, personally attacking, or uncivil language Anderson, Brossard, Scheufele, Xenos, and Ladwig, 2014; Nobata et al., 2016; Wulczyn, Thain, and Dixon, 2017), are not presently available (Warner and Hirschberg, 2012), or rely on keywords (e.g., from HateBase[3]) rather than annotation (Mathew et al., 2019).

Notable partial similarities between our work and previous datasets have to do with shared domain (i.e., Gab), similar operationalization of hate speech definitions, and similar typologies. Several previous datasets have also annotated Gab posts (Mathew et al., 2020; Qian et al., 2019), though one key difference is that these previous works used keywords and lists of common slurs to filter data before annotation. The most similar work in terms of hate speech operationalization are de Gibert et al. (2018) and Vidgen and Yasseri (2020), which both take on the problem of recognizing and categorizing particular hateful

[3] https://hatebase.org/

[4] Many data are unavailable due to IDs having been deleted; see https://github.com/ZeerakW/hatespeech/issues/11

[5] via https://about.fb.com/news/2017/06/hard-questions-hate-speech/

Table 1

English-language hate speech datasets most similar to the present work, mostly drawn from Poletto et al. (2020). Datasets listed here are manually annotated according to a hate speech typology, versus "flagged" as offensive or abusive on the basis of metadata.

Source	Domain	Hate Speech Definition	Hate Labels
Waseem and Hovy (2016)[4]	Twitter	Using slurs, attacking out-groups by employing rhetorical devices	Sexism, Racism
Davidson, Warmsley, Macy, and Weber (2017)	Twitter	Expression of hatred, derogation, humiliation, or insult towards out-group	Hate, Offensive
de Gibert, Perez, García-Pablos, and Cuadros (2018)	Stormfront	Attacks directed at out-group, motivated by out-group's identity	Hate, Hate-relation
Qian et al. (2019)	Reddit & Gab	A direct attack on people based on protected characteristics[5]	Hate
Mathew, Saha, Yimam, Biemann, Goyal, and Mukherjee (2021)	Twitter & Gab	Follows Davidson et al. (2017)	Hate/Offensive, Target, Rationale
Vidgen and Yasseri (2020)	Twitter	Expressing or sharing indiscriminate negativity against Islam or Muslims	Weak Islamaphobia, Strong Islamophia

ideologies. In particular, de Gibert et al. takes on the problem of identifying the expression of white supremacist views, while Vidgen and Yasseri addresses Islamophobic language and beliefs. And in several works listed in Table 1, more informative typologies (i.e., going beyond the binary hate label) have been proposed which are similar to our own. One novelty of note in these works is from Mathew et al. (2020), specifically the use of "rationales" where annotators selected the span of text explaining their labeling.

Though similarities exist between our operationalization, domain, and typology and those of previous works, the most notable difference is our operationalization of the top-level categories. Structurally, our operationalization of hate-based rhetoric as either HD, CV, or both is a departure from previous work. By identifying these two dimensions of hate-based rhetoric, we allow for more nuanced analysis that can better isolate rhetorical texts that convey an incitement to hatred, versus outright calls for violence. More fundamentally, however, our detailed operationalization based on legal precedent and theories and findings from the social sciences results in an altogether different meaning for HD and CV, which again are based on the motivations of those producing hate speech and how they accomplish it through language. And finally, by following the recent, encouraging trend of incorporating targeted populations and framing (Waseem et al., 2017) into hate speech studies, we maximize the usefulness and comprehensiveness of our annotations.

Methodology

Data Collection

Sampling documents for hate speech annotation is difficult due to sparsity and are consequently non-representative (Wiegand, Ruppenhofer, and Kleinbauer, 2019). Since randomly sampling posts from venues like Twitter results in very few hate speech examples, keywords and topic-based sampling strategies are often 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 et al., 2018).

Resolving the data quality issues that come from biased sampling is an open issue.

However, it is less so when the underlying population posts a disproportionately large amount of hate speech, as is the case with Gab and similar platforms, such as the "Stormfront" Reddit community (de Gibert et al., 2018). Thus, randomly sampled posts from such domains yield a more representative dataset, though potentially atypical in other ways given the irregularity of a community that frequently posts hate speech.

We downloaded Gab posts from the public dump of data by Pushshift.io[6] (Gaffney, 2018). 28,000 posts were randomly sampled, stratified so that they were equally balanced across months from the time period of January, 2018 to October, 2018. The size of this sample was roughly supported by previous datasets, for example Davidson et al. (2017). Posts were only considered which had less than four non-hyperlink, non-hashtag tokens.

Annotation Process

Coding manual. A coding manual was generated based on the same sources as discussed above. This document contains the initial description of hate-based rhetoric, including its legal foundations and similarities to typologies in computational research, as well as examples used to instruct annotators on each category. For purposes of transparency and the potential reapplication of our coding method, the original manual, including definitions, discussion of categories, and examples, is included in Supplemental Materials, and the most salient components are reiterated in this section.

The definition of hate-based rhetoric presented to annotators before training (in addition to reading the entire manual) was:

> 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 (p. 8).

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

An illustrative set of examples from Gab that were annotated collectively by the authors, with their labels agreed upon, are shown in Table 2. These examples are representative of the dehumanization and violence of hate-based rhetoric. Not all target populations are represented; for a full description of the resulting target population tags for the annotated dataset, see Section above entitles "Targeted population for HD and CV."

Annotator training and monitoring. 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 by the authors.

A pressing concern in the collection of these data is the potential for annotators to experience trauma or similar negative effects, such as desensitization. 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 (Kwan, Dickson, Richardson, MacDowall, Burchett, Stansfield, Brunton, Sutcliffe, and Thomas, 2020; Levin, 2017; Ybarra, Mitchell, Wolak, and Finkelhor, 2006). Exposure to hate speech is associated with symptoms of trauma exposure, higher liability assessments of targets, and low self-esteem (Boeckmann and 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), and the broader issue of harms to content moderators has been the subject of recent scholarship (Roberts, 2019). 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 and Figley, 2017; May and Wisco, 2016). These consequences have been studied in police officers, first responders, and social workers, among others (Kleim and Westphal, 2011; Perez, Jones, Englert, and Sachau, 2010; Wagaman, Geiger, Shockley, and 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, and Dalton, 2003; Kaplan, Bergman, Christopher, Bowen, and Hunsinger, 2017; Kleim and Westphal, 2011).

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 and Westphal, 2011).

Annotators accessed the raw text of Gab posts via a secure online portal created for text annotation, and could halt at any time. The number of posts annotated per annotator

(n = 18, M = 5,109, Mdn = 4,044) ranged from 288 to 13,543. Posts were discarded that were annotated by less than three annotators.

Label Distributions and Inter-Annotator Agreement

To evaluate inter-annotator agreement, we computed Fleiss's kappa for multiple annotators (Fleiss, 1971) and the Prevalence-Adjusted and Bias-Adjusted Kappa (PABAK; Byrt, Bishop, and Carlin, 1993) for Fleiss. PABAK, which lowers the impact of "expected agreement" relative to standard kappa, can be used with imbalanced labels (e.g., the "CV" label is positive, across annotators, for less than 1 percent of posts) as kappa is known to underestimate annotator agreement in such cases. Fleiss's kappa and PABAK were computed for the three top-level categories — HD, CV, and VO — over the entire dataset. For target population and framing labels, kappa and PABAK were computed only for documents where a majority of annotators judged the text to be either HD or CV. The outcomes of inter-annotator agreement analysis are shown in Table 3.

In Table 3, the binary distributions of each label (majority vote, ties resolved as positive) for three subsets of the corpus are given: the entire corpus, the portion labeled by the majority as HD, and the portion labeled by the majority as CV. Note that the distributions for the latter two subsets are not necessarily exhaustive: in some cases, not all annotators labeled a document as hate-based rhetoric (e.g., as HD). In this scenario, they did not label the target group or the framing label, so a proper majority could not be computed. Thus, each of the targeted populations and framing labels contain a significant subset where no consensus could be drawn from the annotations.

Approximately 9% of the entire corpus is either HD or CV. Roughly one third of all HD documents were judged by a majority as containing vulgar language, while over 40% of calls for violence contained the same. The two most common target populations were race/ethnicity (∼29% of HD and ∼14% of CV) and religious identity (∼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, with only 4% of CV documents doing the same.

Annotating hate speech has been documented to result in high levels of annotator disagreement (e.g., Ross, Rist, Carbonell, Cabrera, Kurowsky, and Wojatzki, 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 et al., 2020). Compounding this issue has been the fact that positive labels suffer sparsity issues. While our annotators are far from identical in their labeling tendencies, sufficient agreement is obtained for top-level categories (HD, CV, and VO) and for target populations. However, the low agreement for Implicit/Explicit rhetoric casts doubts on their usefulness and informativeness. In the future, follow-up work can investigate how to improve these labels.

Data Release

The GHC is available on the Open Science Framework (OSF)[7]. Included in the release is a file containing every annotation used in the GHC, designated *train* and *test* majority-aggregated datasets for future analysis, and the original document used to train annotators.

Annotator background data was collected in order to provide future works an opportunity to analyze the relationship between annotator characteristics and annotating behavior on this task. Implicit Association Test measures (Greenwald, McGhee, and Schwartz, 1998) and measures on attitudes towards hate crime and hate speech censorship and policies (Cabeldue, Cramer, Kehn, Crosby, and Anastasi, 2018) were collected for 8 of the annotators, covering approximately 62% of the annotations in total. In the future, we will also release accompanying annotations for the GHC posts which were performed

[7] `https://osf.io/edua3/`

during this process with respect to the moral concerns typology used by Hoover, Portillo-Wightman, Yeh, Havaldar, Davani, Lin, Kennedy, Atari, Kamel, Mendlen, Moreno, Park, Chang, Chin, Leong, Leung, Mirinjian, and Dehghani (2020).

Experiments

A series of classification experiments, accompanied by a qualitative examination of feature importance for the trained classifiers, was performed. This was undertaken in order to establish performance baselines for modeling with the GHC as well as to give qualitative assessments on the content of hate-based rhetoric categories.

Below, performance is reported for a single, held-out test set of posts. This test set is made public for replicability and comparison purposes.

Classification of Top-Level Categories

One of the objectives of the GHC is to increase the ability to train robust classification algorithms for hate-based rhetoric, which serve a number of purposes. Classifiers, once trained on these data, can be applied to new data in order to estimate hate-based rhetoric labels. Here we test the classification performance of three standard methods on the HD, CV, and VO labels.

Methods. First, feature-based models were constructed with features generated via the 2015 release of the Linguistic Inquiry and Word Count dictionary (LIWC; Pennebaker, Boyd, Jordan, and Blackburn, 2015) and via Term Frequency-Inverse Document Frequency (TF-IDF) weights (Aizawa, 2003; Jones, 1972). By computing the frequency of LIWC word categories (each LIWC feature was also scaled by the maximum value in the train set), posts are represented by a 73-length vector corresponding to psychologically or grammatically meaningful types of words. Here, LIWC is used to benchmark the performance of lexicon approaches in addition to the ease of interpreting LIWC features in post hoc feature importance analysis. Meanwhile, TF-IDF vectors are a strong baseline with similar interpretability benefits (see Joulin, Grave, Bojanowski, and Mikolov, 2017).

To measure whether there is complementary information between TF-IDF and LIWC feature representations, an additional feature set was generated by concatenating TF-IDF weights and LIWC counts.

Support Vector Machines (SVMs; Cortes and Vapnik, 1995; Joachims, 1998) with linear kernels were implemented for each feature set. SVM models were fit to TF-IDF and LIWC feature sets with the *scikit-learn* v0.22.1 (Pedregosa, Varoquaux, Gramfort, Michel, Thirion, Grisel, Blondel, Prettenhofer, Weiss, Dubourg, Vanderplas, Passos, Cournapeau, Brucher, Perrot, and Duchesnay, 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 that C (controlling level of regularization), *penalty* (L-1, promoting sparsity, or L-2, the 'Ridge' penalty), the *loss* (squared hinge loss or hinge loss), and the formulation of the optimization problem ('Dual' or 'Primal') were selected based on 5-fold cross validated grid search on the training set. For all classifiers, *class weight* was set to 'balanced' (assigning higher importance to predicting positive classes, proportional to class distribution in the training set).

Language model fine-tuning. We applied language model fine-tuning, which is a powerful technique for transfer learning in NLP and consists of the transfer of learned linguistic knowledge from neural language models to potentially small-scale downstream applications (Devlin, Chang, Lee, and Toutanova, 2018; Howard and Ruder, 2018; Liu, Gardner, Belinkov, Peters, and Smith, 2019; Peters, Neumann, Iyyer, Gardner, Clark, Lee, and Zettlemoyer, 2018; Radford, Narasimhan, Salimans, and Sutskever, 2018). We used one of the most successful techniques, fine-tuning "Bidirectional Encoder Representations from Transformers" models (BERT; Devlin et al., 2018), pretrained on large text corpora. BERT consists of a large neural network (either 12 or 24 layers) which learns an advanced degree of compositionality, syntax, and word meaning through the tasks of predicting missing words in sentences as well as the order of sentences. The efficacy of BERT and similar models is in transfer: after training on massive corpora of books, online media, and

Wikipedia, models are saved, distributed, and subsequently “fine-tuned” on downstream tasks. We adopted this fine-tuning approach, in which a pre-trained BERT model was applied to a given prediction task and model weights were adjusted in order to maximize classification performance. BERT models were fine-tuned following the instructions of (BERT; Devlin et al., 2018) using the transformers [8] (Wolf, Debut, Sanh, Chaumond, Delangue, Moi, Cistac, Rault, Louf, Funtowicz, and Brew, 2019) Python library in PyTorch (Paszke, Gross, Massa, Lerer, Bradbury, Chanan, Killeen, Lin, Gimelshein, Antiga, Desmaison, Köpf, Yang, DeVito, Raison, Tejani, Chilamkurthy, Steiner, Fang, Bai, and Chintala, 2019).

Of note in the design of classifiers for our dataset is the issue of class imbalance. In Table 3, we reported that less than 10% of instances are either HD or CV. To address this, we applied resampling, which weighs instances by their class proportion (learned from the distribution in the training set)[9]. This approach is similar to oversampling of the positive (under-represented) class, which is common in hate speech classification research (see Arango, Péerez, and Poblete, 2019), but is less likely to suffer from overfitting due to duplicate positive instances during training. Future works can attempt to further mitigate the issues of class imbalance by using dedicated techniques (e.g., Ah-Pine and Soriano-Morales, 2016).

Hyperparameter tuning was performed via the Optuna library (Akiba, Sano, Yanase, Ohta, and Koyama, 2019), where learning rate (uniform distribution over 5e^{-5} to 5e^{5}), weight decay (uniform over 0.0 to 0.5), and warmup steps (uniform over 0 to 500) were tuned by sampling over parameter values, selecting the best performing trial on the validation set. Training batch size was consistent across models at 16. Model checkpoints were saved and evaluated every 500 steps (i.e., batches), with the best-performing checkpoint returned. For each label, 10 trials were performed with each trial taking a

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

[9] See `https://github.com/ufoym/imbalanced-dataset-sampler`

sample of hyperparameters, and the model checkpoint corresponding to the best trial (F_1) was reported. Training took place on a single Nvidia GeForce GTX 1080 GPU, with each training trial taking approximately 25 minutes (we note that training took longer due to resampling; without resampling each epoch took less than 5 minutes).

Results. In Table 4, precision, recall, and F_1 score for the positive class on the test set are reported for predicting HD, CV, and VO. F_1 score is the harmonic mean between precision and recall and is commonly used to evaluate classification on imbalanced data, as high accuracy can be trivially maximized by predicting no-hate for all instances. In addition, performance metrics are given for two random predictions: "Random (Balanced)" denotes comparison of true test labels with a sample from a Bernoulli distribution with $p = 0.5$; "Random (Weighted)" is identical in nature, excepting that p is selected as the proportion of positive instances in the training set. All models were selected based on validation F_1, with fine-tuned models evaluated on a single, independent dataset (20% of the remaining train set) and SVM models evaluated on 5-fold cross-validation on the train set (80% of entire dataset). For fine-tuned models, hyperparameters were selected based on randomized trials using the *optuna* library in Python (Akiba et al., 2019). For SVM models, hyperparameters were selected with grid search.

All models out-performed random choice, showing that even LIWC features are able to meaningfully predict each label. Additionally, the low performance observed is contextualized by the skewed distribution of the labels, which makes high F_1 difficult to obtain.

The highest average performance in all three cases was obtained by fine-tuning the pre-trained BERT model. The scores obtained via LIWC represent the extent to which expert-defined word categories, not adapted to the domain, are able to describe the occurrence of hate-based rhetoric. Their performance indicates that LIWC categories capture relatively less information than TF-IDF for HD, CV, and VO, yet improve performance when added to the TF-IDF features for HD and CV. We note that, due to the fact that VO is essentially a lexical category describing profanity and slurs, it is to be expected that TF-IDF performs well, and that LIWC adds comparatively little information to TF-IDF. And congruently to leading NLP research into fine-tuning language models for text classification, the superior performance of fine-tuning BERT indicates that hate-based rhetoric is defined by the composition of words into more sophisticated meaning representations.

These classification metrics should be interpreted as baseline measures which inform and motivate future modeling. For example, our results show that lexicon-based approaches such as LIWC are particularly ill-suited to application on our dataset. Actual application of GHC-trained classifiers to new data will likely require multiple iterations of modeling improvements, and potentially new approaches to the problem (e.g., including background knowledge into classifiers).

As noted by Arango et al. (2019), while many works on hate speech have reported high, application-ready performance metrics, the reality is that hate speech prediction is an extremely difficult task, and F_1 statistics on in-domain test datasets falls in a similar range as observed in our experiments. Specifically, Arango et al. (2019) found that true F_1 performance on the Twitter dataset from Waseem and Hovy (2016) is below 50%, similar to our results with fine-tuning BERT to the HD label. In addition, performance in these experiments are in line with previous hate speech classification for non-trivial predictions tasks in (MacAvaney, Yao, Yang, Russell, Goharian and Frider, 2019). Of note in this work, BERT models achieved 0.52 F_1 on an “aggression detection” dataset (Kumar, Ojha, Malmasi, and Zampieri, 2018), while other tasks observed much higher performance (e.g., F_1 of 0.89 on Twitter hate speech Davidson et al., 2017).

Feature Importance Analysis

Feature importance was analyzed via model coefficients for the TF-IDF and LIWC SVM models. The words and LIWC categories with the highest importance for classifying HD, CV, and VO are shown in Table 5.

In considering the most important features in classifying Human Degradation, Calls for Violence, and Vulgarity/Offensive Language, one can see clearly the themes of these categories in the GHC. In the case of Calls for Violence, violent and aggressive verbs (e.g., “kill”, “hang”, or “nuke”) are prominent, while in the case of Vulgar/Offensive there is an obvious presence of slurs. In the case of Human Degradation, the most predictive words are predominately references to social groups — what others have called “Social Group Terms” (Dixon et al., 2018) — for example “white,” “Muslims,” and “Jews.”

In contrast to TF-IDF features, which produce fine-grained insights into the vocabulary of hate-based rhetoric, LIWC features give a sense of the aggregate linguistic trends. Table 6 shows the top 5 LIWC categories by feature weight in SVMs.

Misclassification Analysis

For the application of classifiers trained on the GHC, higher performance is required relative to the F_1 levels observed in our classification experiment. Particularly for automatic content moderation, recall and precision scores of 60% and 42%, respectively, are well-below the needed threshold. Additional work is required in order to understand and correct model mistakes. To help the development of improved approaches to classifying posts in the GHC, here we attempt to diagnosis sources of misclassification.

Table 7 shows examples of different types of misclassification errors observed in the test set. Examples are grouped into false positive errors, false negative errors, and general sources of error which affect both precision and recall. We identified two patterns of GHC posts which were commonly misclassified as “positive” when their true label was negative for both HD and CV. First, non-hateful usage of moral rhetoric, insulting language (e.g., profanity), non-hateful stereotyping, and conspiracy theorizing was incorrectly predicted as hate-based rhetoric. For example, the sentence “Because black dudes don’t play video games” was a false positive prediction which, because it’s stereotyping was not harmful or dehumanizing, had a true label of non-hateful. Similarly, many false positives contained mentions of social group terms — e.g., “black,” “jewish,” or “transgender.” Together, these two issues accounted for the majority of the false positive errors.

Low recall can largely be explained by models’ failure to understand rhetorical sentences, which is one of the central challenges posed by the GHC to researchers (see Table 7). For example, “Germany’s stuck with the roaches they invited in” is an anti-immigrant statement comparing people to insects, but a model will only be able to recognize such a complex comparison if the comparison is contained in the training data, which evidently was not the case.

Lastly, general sources of error include limitations of our typology (i.e., ambiguous posts) or incorrect annotator labels. Ambiguous posts were typically offenses that possessed some, but not all of the characteristics of hate-based rhetoric. For example, degrading language used toward more powerful or advantaged groups (e.g., “[Conservatives] lack intelligence” in Table 7) was not thoroughly discussed in our coding typology, thus annotators varied in their assessment of the harms of such posts. Incorrect labels can be attributed to noise in the annotation process, and improvements on the GHC will correct mislabeling.

Conclusion and Future Work

We introduced the GHC, a large-scale corpus annotated for hate-based rhetoric. In doing so, we operationalized legal standards regarding the incitement to hatred in countries such as Germany and theoretical conceptions of prejudice and hateful ideology in the social sciences. The presence of human degradation and calls for violence were our primary labels, as well as more fine-grained categories — e.g., targeted populations, vulgarity, and framing.

As a practical tool, the GHC can be used in NLP research, specifically with regard to hate speech detection, and in data-driven studies of hatred and prejudice in the social sciences.

The GHC 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.

Practical Applications

There are two main categories of research which can apply the GHC. First, researchers can use the corpus in order to train and evaluate hate speech classifiers. Here, the GHC complements existing datasets, with its characteristics making it ideal for detecting rhetoric-heavy hate speech, versus slurs and explicit harms.

Other than being used for achieving automated content moderation, classifiers can be trained on the GHC and subsequently applied to new data in order to analyze hate speech trends across large, representative samples of data (e.g., social media posts). For example, in order to quantify the relationship between "moral sentiment" (the framing or presentation of an object as moral or possessing moral value), Hoover, Atari, Davani, Kennedy, Portillo-Wightman, Yeh, Kogon, and Dehghani (2019) trained hate speech classifiers on a subset of the GHC and applied them to more than 20 million Gab posts, thereby analyzing the correlation of moral and hateful sentiments across an exhaustive sample.

As discussed in the Introduction, the operationalization of hate speech used to annotate a given corpus informs the ways that model predictions, having been trained using that corpus, can be interpreted. Since hate-based rhetoric draws on the ideological and prejudicial roots that lead to hate speech — moreso than other annotated datasets — studies such as Hoover et al. (2019) are better able to extrapolate relationships between hate more generally and other phenomena.

Broader Implications and Future Directions

One assumption that informed the operationalization of hate-based rhetoric is that recognizing speaker intentions is key to recognizing hate speech. Specifically, we believe that recognizing hate speech is at least partially dependent on recognizing the communicative intentions of the speaker. In this case, this amounts to recognizing the goals of hate speech, which are to dehumanize, disempower, and otherwise work to advance prejudicial ideology against marginalized social groups. Of course, this complicates the task of annotation, as observed in multiple studies of annotator disagreement for hate speech (Mostafazadeh Davani et al., 2020; Ross et al., 2017), and raises questions about the feasibility of automated content moderation for hate speech based on the GHC. However, the alternative is to ignore the fact that prejudices are often delivered implicitly and are perhaps most powerful and harmful in this content than explicit prejudice (e.g., racial slurs) that can easily be detected and removed.

One implication of this view is that, in order to make progress on understanding and predicting implicit hate speech, more extra-linguistic data ought to be used in models of hate speech. The same text, coming from difference venues or user accounts, might mean different things with regards to communicated prejudices (e.g., the concept of a "dog-whistle" to signal racist policies or opinions to supporters; López, 2015). We reiterate that greater collaborations between NLP researchers and social scientists can help to develop datasets and modeling strategies that are sensitive to speaker intentions.

Remaining Challenges

Despite our intentions to mitigate the problem of imbalanced datasets in constructing the GHC from a random sample, recent work has found that models trained on the GHC disproportionately associate the presence of group-identifying terms (e.g., "black" and "jewish") with hate-based rhetoric (Kennedy, Jin, Mostafazadeh Davani, Dehghani, and Ren, 2020). Thus, accounting for these biases is a major task for not only those working with the GHC, but also those analyzing other hate speech datasets (see Dixon et al., 2018; Sap, Card, Gabriel, Choi, and Smith, 2019; Waseem and Hovy, 2016). One potential solution was proposed by Kennedy et al., which focused on classification of hate speech in the GHC as well as a corpus of posts from the white supremacist forum “Stormfront” (de Gibert et al., 2018). In this work, the authors used post hoc explanations (Jin, Du, Wei, Xue, and Ren, 2019) to determine which phrases in a given sentence influenced positive predictions. Then, with these word- and phrase-based importance scores, models were regularized to place less importance on social group terms.

Improvements to the GHC can be undertaken in order to further resolve inter-annotator disagreements; understand and mitigate model biases that emerge as a result of “annotator stereotyping,” or differential treatment of particular social groups by annotators on account of existing social stereotypes (Mostafazadeh Davani et al., 2020); and refinements to the annotation process itself in order to minimize exposure of annotators to harmful content — e.g., by combining computational and manual approaches to labeling. These improvements can help to improve the GHC as a resource, both in terms of the correctness of annotations and the clarity of our coding typology.

Lastly, we note that our approach, like many previous works in annotating hate speech, assume a Euro-centric or “Western” worldview (i.e., we consider the hate speech laws of the United States, Canada, the Netherlands, and Germany). This obviously biases our work in key, perhaps imperceptible ways. In order to avoid the “home-field disadvantage” (Medin, Bennis, and Chandler, 2010) by assuming the Western conception of hate speech (and the debates on free speech), those who use this work ought to recognize this as assumed, rather than as the default. In addition, future work can repeat the steps we took in operationalizing hate-based rhetoric — reviewing legal literature and scientific research — for non-Western countries, languages, and cultures.

Conflict of interest

The authors declare that they have no conflict of interest.

Table 2

Annotated examples from Gab illustrating the hate-based rhetoric typology.

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
#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
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
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

Table 3

Inter-annotator agreement kappas and binary label distributions (percent positive). For targeted populations and framing labels, inter-annotator agreement kappas and distributions are computed for the subset of labels which a majority of annotators labeled as either HD or CV. Distributions (last three columns) are given for the overall corpus, the subset that are majority-labeled as HD, and the subset that are majority-labeled as CV.

Tag	Description	Fleiss (PABAK)	Label Distributions		
			All	HD	CV
		Top-Level Categories (N=27,655)			
HD	Human Degradation	0.23 (0.67)	8.5%	100.0%	34.2%
CV	Calls for Violence	0.28 (0.97)	0.6%	2.3%	100.0%
VO	Vulgar/Offensive	0.30 (0.79)	6.3%	33.3%	41.3%
		Targeted Populations & Framing			
REL	Religious Identity	0.52 (0.69)	3.1%	21.9%	18.7%
RAE	Racial/Ethnic Identity	0.48 (0.59)	3.5%	29.4%	13.5%
SXO	Sexual Orientation	0.53 (0.90)	0.7%	5.9%	3.9%
GEN	Gender Identity	0.47 (0.87)	0.9%	7.1%	1.9%
IDL	Ideology	0.29 (0.66)	1.6%	10.5%	14.8%
NAT	Nationality	0.26 (0.69)	1.5%	9.6%	10.3%
POL	Political Identity	0.25 (0.61)	2.3%	14.5%	12.3%
MPH	Mental/Physical Health	0.18 (0.92)	0.2%	1.6%	1.3%
IM	Implicit Rhetoric	0.03 (0.38)	1.8%	14.1%	3.9%

Table 4

F_1, precision, and recall for predicting HD, CV, and VO in the held-out test set.

	Method	**F_1**	**Recall**	**Precision**	**Accuracy**
Human Degradation	Random (Balanced)	15.4	50.0	9.1	50.7
	Random (Weighted)	8.4	8.4	8.4	83.7
	SVM_{LIWC}	33.3	56.2	23.7	80.3
	$\text{SVM}_{\text{TF-IDF}}$	44	54.1	37	87.9
	$\text{SVM}_{\text{TF-IDF+LIWC}}$	45.2	66.5	34.2	85.9
	$\text{FineTune}_{\text{BERT}}$	48.9	59.7	41.5	89.1
Calls for Violence	Random (Balanced)	0.8	45.8	0.4	49.8
	Random (Weighted)	0.0	0.0	0.0	98.9
	SVM_{LIWC}	2.8	50	1.4	82.2
	$\text{SVM}_{\text{TF-IDF}}$	13.9	25	9.6	98.4
	$\text{SVM}_{\text{TF-IDF+LIWC}}$	14.5	32.1	9.4	98.1
	$\text{FineTune}_{\text{BERT}}$	32.3	35.7	29.4	99.2
Vulgar or Offensive	Random (Balanced)	11.9	50.7	6.8	49.9
	Random (Weighted)	7.2	7.0	7.4	87.9
	SVM_{LIWC}	37.3	46.1	31.2	90.6
	$\text{SVM}_{\text{TF-IDF}}$	48.1	56	42.2	92.7
	$\text{SVM}_{\text{TF-IDF+LIWC}}$	45.3	65.8	34.5	90.3
	$\text{FineTune}_{\text{BERT}}$	51.4	55.1	48.2	93.7

Table 5

The top 10 features by coefficient value in linear SVMs with TF-IDF features.

Human Degradation	Calls for Violence	Vulgar/Offensive
jews (3.46)	shoot (1.98)	fuck (4.76)
white (3.33)	shot (1.95)	shit (4.31)
jew (2.94)	kill (1.75)	fucking (3.75)
n****r (2.62)	hang (1.58)	ass (3.16)
faggot (2.59)	mil (1.56)	bitch (3.14)
muslim (2.51)	throw (1.49)	faggot (2.92)
muslims (2.42)	simply (1.47)	n****rs (2.79)
islam (2.42)	communists (1.47)	n****r (2.73)
n****s (2.30)	nuke (1.43)	jews (2.32)
fucking (2.26)	nailed (1.42)	shithole (2.20)

Table 6

The top 5 LIWC categories by coefficient value in linear SVMs. Coefficient values are given in parentheses, and are only meaningful in a relative sense. Only one LIWC category was a non-zero predictor for Vulgar/Offensive, as the regularization selected for this model during grid search (L-1 regularization) produced sparse coefficients.

Human Degradation	Calls for Violence	Vulgar/Offensive
Religion (8.73)	Discrepancy (5.85)	Swear words (1.54)
Swear words (6.00)	Swear words (3.38)	
Anger (3.74)	3^{rd}/plural Pronouns (2.86)	
Sexual (3.50)	Religion (2.27)	
3^{rd}/plural Pronouns (3.07)	Anger (2.16)	

Sources of False Positive Errors	
Non-hateful moral rhetoric, insults, and accusations	President Trump can't solve the problem of high health insurance costs with these *traitors* blocking his efforts.
	Obama is a *filthy liar*
	Yes I do and they almost did a fantastic job you *racist*, give the people some credit.
Group terms in non-hate contexts	A *#transgender woman* in #Iran was abused, but the police refused to respond, blaming the abuse on how she looked.
	I LOVE this amazingly smart, courageous, beautiful *black* lady and ALL *black* people (and all races for that matter) should raise her up on a pedestal and consider her their voice for reason and sanity and betterment!!!
Sources of False Negative Errors	
Unrecognized rhetoric	The British people are at war with the invading third world
	Tranvesties are just as fucked up as pedos. Both should be banned from all western societies.
	Yeah... Germany's stuck with the roaches they invited in. Ha!
Unrecognized threats and calls for violence	I think the title is "Gas the democrats, political war now!"
	These filthy bastards are all connected. One begins to understand the impetus for total purges.
	No pension for traitors....just lethal injections!
General Sources of Errors	
Ambiguous with respect to our typology	…I say that if your religion requires you to kill on its behalf then it's not a religion at all. It's a DEATH CULT
	With conservatives? yes. they lack intelligence.
	Real friends don't let their friends get infected with communism
Annotator Error	His future is defo in porn since he can't spell
	Agreed, but I think a lot of the Anti Semitism is democrat plants trying to make Gab out to look like a right-wing extremist site.

Table 7

Types of misclassification errors, with examples for each. Emphasis (italics) are added to indicate the portion of the text relevant to each misclassification category.

References

Ah-Pine J, Soriano-Morales, EP (2016). A study of synthetic oversampling for twitter imbalanced sentiment analysis. In Workshop on interactions between data mining and natural language processing (DMNLP 2016).

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

Akiba T, Sano S, Yanase T, Ohta T, Koyama M (2019) Optuna: A next-generation hyperparameter optimization framework. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp 2623–2631

Akiwowo S, Vidgen B, Prabhakaran V, Waseem Z (eds) (2020) Proceedings of the Fourth Workshop on Online Abuse and Harms, Association for Computational Linguistics, Online, URL `https://www.aclweb.org/anthology/2020.alw-1.0`

Allport GW (1954) The Nature of Prejudice. Addison-Wesley

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

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

Andrews TM (2021) Gab, the social network that has welcomed qanon and extremist figures, explained. `https://www.washingtonpost.com/technology/2021/01/11/gab-social-network/`, accessed: 2021-02-15

Atari M, Davani AM, Kogon D, Kennedy B, Saxena NA, Anderson I, Dehghani M (2021) Morally homogeneous networks and radicalism

Arango, A, Pérez, J, Poblete, B (2019) Hate speech detection is not as easy as you may think: A closer look at model validation. In Proceedings of the 42^{nd} International ACM SIGIR Conference on Research and Development in Information Retrieval, 45–54

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

Boeckmann RJ, 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 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

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

Cabeldue MK, Cramer RJ, Kehn A, Crosby JW, Anastasi JS (2018) Measuring attitudes about hate: Development of the hate crime beliefs scale. Journal of interpersonal violence 33(23):3656–3685

Chaplinsky v New Hampshire (1942) Chaplinsky v. New Hampshire

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. In: Eleventh international aaai conference on web and social media

Devlin J, Chang MW, Lee K, Toutanova K (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:181004805

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, ACM, pp 67–73

Facebook (2020) Community Standards: 12 Hate Speech. https://www.facebook.com/communitystandards/hate_speech, accessed: 2020-02-07

Fan RE, Chang KW, Hsieh CJ, Wang XR, Lin CJ (2008) Liblinear: A library for large linear classification. Journal of machine learning research 9(Aug):1871–1874

Fleiss JL (1971) Measuring nominal scale agreement among many raters. Psychological bulletin 76(5):378

GAB AI, INC (2020) Website terms of service. https://gab.com/about/tos, accessed: 2021-02-15

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

German Criminal Code (1998) German Criminal Code. https://www.gesetze-im-internet.de/englisch_stgb/englisch_stgb.html

de Gibert O, Perez N, García-Pablos A, Cuadros M (2018) Hate speech dataset from a white supremacy forum. In: Proceedings of the 2nd Workshop on Abusive Language Online (ALW2), pp 11–20

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

Google (2020) Hate speech policy. https://support.google.com/youtube/answer/2801939, accessed: 2020-02-07

Greenwald AG, McGhee DE, Schwartz JL (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 URL https://www.wired.com/2016/09/gab-alt-rights-twitter-ultimate-filter-bubble/

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

Hoover J, Atari M, Davani AM, Kennedy B, Portillo-Wightman G, Yeh L, Kogon D, Dehghani M (2019) Bound in hatred: The role of group-based morality in acts of hate. PsyArxiv Preprint 1031234/osfio/359me

Hoover J, Portillo-Wightman G, Yeh L, Havaldar S, Davani AM, Lin Y, Kennedy B, Atari M, Kamel Z, Mendlen M, Moreno G, Park C, Chang TE, Chin J, Leong C, Leung JY, Mirinjian A, Dehghani M (2020) Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment. Social Psychological and Personality Science

Howard J, Ruder S (2018) Universal language model fine-tuning for text classification. arXiv preprint arXiv:180106146

Howard JW (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:191106194

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

Jones KS (1972) A statistical interpretation of term specificity and its application in retrieval. Journal of documentation

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 JB, Bergman AL, 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. In: Proceedings of the 2020 Annual Conference of the Association for Computational Linguistics

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

Kuklinski JH, Cobb MD, Gilens M (1997) Racial attitudes and the "new south". The Journal of Politics 59(2):323–349

Kumar, R and O, AK and Malmasi, S and Zampieri, M (2018) Benchmarking Aggression Identification in Social Media. In Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018): 1–11.

Kwan I, Dickson K, Richardson M, MacDowall W, Burchett H, Stansfield C, Brunton G, Sutcliffe K, Thomas J (2020) Cyberbullying and children and young people's mental health: a systematic map of systematic reviews. Cyberpsychology, Behavior, and Social Networking 23(2):72–82

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 URL https://www.theguardian.com/technology/2017/jan/11/microsoft-employees-child-abuse-lawsuit-ptsd

Liu NF, Gardner M, Belinkov Y, Peters M, Smith NA (2019) Linguistic knowledge and transferability of contextual representations. arXiv preprint arXiv:190308855

López IH (2015) Dog whistle politics: How coded racial appeals have reinvented racism and wrecked the middle class. Oxford University Press

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

MacAvaney, S and Yao, HR and Yang, E and Russell, K and Goharian, N and Frieder, O (2019) Hate Speech Detection: Challenges and solutions. PLoS One 14(8):e0221152.

Mathew B, Dutt R, Goyal P, Mukherjee A (2019) Spread of hate speech in online social media. In: Proceedings of the 10th ACM conference on web science, pp 173–182

Mathew B, Illendula A, Saha P, Sarkar S, Goyal P, Mukherjee A (2020) Hate begets hate: A temporal study of hate speech. Proceedings of the ACM on Human-Computer Interaction 4(CSCW2):1–24

Mathew B, Saha P, Yimam SM, Biemann C, Goyal P, Mukherjee A (2021) Hatexplain: A benchmark dataset for explainable hate speech detection. In: Proceedings of The Thirty-Fifth AAAI Conference on Artificial Intelligence (to appear)

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

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

May CL, Wisco BE (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

Medin D, Bennis W, Chandler M (2010) Culture and the home-field disadvantage. Perspectives on Psychological Science 5(6):708–713

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

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 (in press). In: Proceedings of the 42nd Annual Conference of the Cognitive Science Society (CogSci)

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, International World Wide Web Conferences Steering Committee, pp 145–153

Olteanu A, Castillo C, Boy J, Varshney KR (2018) The effect of extremist violence on hateful speech online. In: Twelfth International AAAI Conference on Web and Social Media

Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, Desmaison A, Köpf A, Yang E, DeVito Z, Raison M, Tejani A, Chilamkurthy S, Steiner B, Fang L, Bai J, Chintala S (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, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12:2825–2830

Pennebaker JW, Boyd RL, Jordan K, Blackburn K (2015) The development and psychometric properties of liwc2015. Tech. rep.

Perez LM, Jones J, Englert DR, 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 ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L (2018) Deep contextualized word representations. arXiv preprint arXiv:180205365

Poletto F, Basile V, Sanguinetti M, Bosco C, Patti V (2020) Resources and benchmark corpora for hate speech detection: a systematic review. Language Resources and Evaluation pp 1–47

Qian J, Bethke A, Liu Y, Belding E, Wang WY (2019) A benchmark dataset for learning to intervene in online hate speech. In: Proceedings of the 2019 Conference on

Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp 4757–4766

Radford A, Narasimhan K, Salimans T, Sutskever I (2018) Improving language understanding by generative pre-training, preprint available at https://www.cs.ubc.ca/~amuham01/LING530/papers/radford2018improving.pdf

RAV v St Paul (1992) Rav v. st. paul

Roberts ST (2019) Behind the screen: Content moderation in the shadows of social media. Yale University Press

Romero L (2021) Experts say echo chambers from apps like parler and gab contributed to attack on capitol. ABC News URL https://abcnews.go.com/US/experts-echo-chambers-apps-parler-gab-contributed-attack/story?id=75141014

Roose K (2018) On gab, an extremist-friendly site, pittsburgh shooting suspect aired his hatred in full. The New York Times URL 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. In: Proceedings of the 3rd Workshop on Natural Language Processing for Computer-Mediated Communication

Sap M, Card D, Gabriel S, Choi Y, Smith NA (2019) The risk of racial bias in hate speech detection. In: Proceedings of the 57th annual meeting of the association for computational linguistics, pp 1668–1678

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)

Twitter (2020) Hateful Conduct Policy.

https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy, accessed: 2020-02-07

Vidgen B, Yasseri T (2020) Detecting weak and strong islamophobic hate speech on social media. Journal of Information Technology & Politics 17(1):66–78

Wagaman MA, Geiger JM, Shockley C, Segal EA (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, Association for Computational Linguistics, pp 19–26

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

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

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, Cistac P, Rault T, Louf R, Funtowicz M, Brew J (2019) Transformers: State-of-the-art natural language processing. arXiv preprint arXiv:191003771

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, International World Wide Web Conferences Steering Committee, pp 1391–1399

Ybarra ML, Mitchell KJ, 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