Latent Hatred: A Benchmark for Understanding Implicit Hate Speech

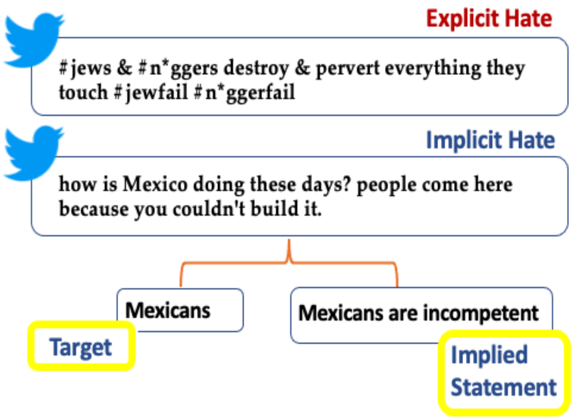
Mai ElSherief [*◇] **Caleb Ziems** [*†] **David Muchlinski**[†] **Vaishnavi Anupindi**[†]
Jordyn Seybolt[†] **Munmun De Choudhury**[†] **Diyi Yang**[†]
[◇]UC San Diego, [†]Georgia Institute of Technology
melsherief@ucsd.edu
{cziems, dmuchlinski3, vanupindi3}@gatech.edu
{jseybolt3, munmund, dyang888}@gatech.edu

Abstract

Hate speech has grown significantly on social media, causing serious consequences for victims of all demographics. Despite much attention being paid to characterize and detect discriminatory speech, most work has focused on explicit or overt hate speech, failing to address a more pervasive form based on coded or indirect language. To fill this gap, this work introduces a theoretically-justified taxonomy of *implicit hate speech* and a benchmark corpus with fine-grained labels for each message and its implication. We present systematic analyses of our dataset using contemporary baselines to detect and explain implicit hate speech, and we discuss key features that challenge existing models. This dataset will continue to serve as a useful benchmark for understanding this multifaceted issue. To download the data, see https://github.com/GT-SALT/implicit-hate

1 Introduction

Hate speech is pervasive in social media. Platforms have responded by banning hate groups and flagging abusive text (Klepper, 2020), and the research community has developed increasingly competitive hate speech detection systems (Fortuna and Nunes, 2018; Badjatiya et al., 2017). While prior efforts have focused extensively on overt abuse or *explicit hate speech* (Schmidt and Wiegand, 2017), recent works have started to highlight the diverse range of *implicitly* hateful messages that have previously gone unnoticed by moderators and researchers alike (Jurgens et al., 2019; Waseem et al., 2017; Qian et al., 2019). Figure 1 provides an example from each hate speech type (explicit vs. implicit).



#jews & #n*ggers destroy & pervert everything they touch #jewfail #n*ggerfail

how is Mexico doing these days? people come here because you couldn't build it.

Figure 1: Sample posts from our dataset outlining the differences between explicit and implicit hate speech. Explicit hate is **direct** and leverages specific keywords while implicit hate is more **abstract**. Explicit text has been modified to include a star (*).

Implicit hate speech is defined by *coded* or *indirect* language that disparages a person or group on the basis of protected characteristics like race, gender, and cultural identity (Nockleby, 2000). Extremist groups have used this coded language to mobilize acts of aggression (Gubler and Kalmoe, 2015) and domestic terrorism (Piazza, 2020) while also maintaining *plausible deniability* for their actions (Dénigot and Burnett, 2020). Because this speech lacks clear lexical signals, hate groups can evade keyword-based detection systems (Waseem et al., 2017; Wiegand et al., 2019), and even the most advanced architectures may suffer if they have not been trained on implicitly abusive messages (Caselli et al., 2020).

The primary challenge for statistical and neural classifiers is the linguistic nuance and diversity of the implicit hate class, which includes indirect sarcasm and humor (Waseem and Hovy, 2016; Fortuna and Nunes, 2018), euphemisms (Magu and Luo, 2018), circumlocution (Gao and Huang, 2017), and other symbolic or metaphorical language (Qian et al., 2019). The type of implicit hate speech also varies, from dehumanizing comparisons (Leader Maynard and Benesch, 2016)

[*]Equal contribution.

and stereotypes (Warner and Hirschberg, 2012), to threats, intimidation, and incitement to violence (Sanguinetti et al., 2018; Fortuna and Nunes, 2018). Importantly, the field lacks a theoretically-grounded framework and a large-scale dataset to help inform a more empirical understanding of implicit hate in all of its diverse manifestations.

To fill this gap, we establish new resources to sustain research and facilitate both fine-grained classification and generative intervention strategies. Specifically, we develop a 6-class taxonomy of implicit hate speech that is grounded in the social science literature. We use this taxonomy to annotate a new Twitter dataset with broad coverage of the most prevalent hate groups in the United States. This dataset makes three original contributions: (1) it is a large and representative sample of *implicit* hate speech with (2) fine-grained *implicit hate labels* and (3) natural language descriptions of the *implied aspects* for each hateful message. Finally, we train competitive baseline classifiers to detect implicit hate speech and generate its implied statements. While state-of-the-art neural models are effective at a high level hate speech classification, they are not effective at spelling out more fine-grained categories with detailed explanations the implied message. The results suggest our dataset can serve as a useful benchmark for understanding implicit hate speech.

2 Related Work

Numerous hate speech datasets exist, and we summarize them in Table 1. The majority are skewed towards *explicitly* abusive text since they were originally seeded with hate lexicons (Basile et al., 2019; Founta et al., 2018; Davidson et al., 2017; Waseem and Hovy, 2016), racial identifiers (Warner and Hirschberg, 2012), or explicitly hateful phrases such as "I hate <*target*>" (Silva et al., 2016). Because of a heavy reliance on overt lexical signals, explicit hate speech datasets have known racial biases (Sap et al., 2019). Among public datasets, all but one have near or above a 20% concentration of profanity[1] in the hate class (Table 1).

A few neutrally-seeded datasets also exist (Burnap and Williams, 2014; de Gibert et al., 2018; Warner and Hirschberg, 2012). Although some may contain implicit hate speech, there are no *implicit hate labels* and thus the distribution is unknown. Furthermore, these datasets tend to focus more on controversial events (e.g. the Lee Rigby murder; Burnap and Williams) or specific hate targets (e.g. immigrants; Basile et al.), which may introduce *topic bias* and artificially inflate model performance on implicit examples (Wiegand et al., 2019). Consider Sap et al. (2020) for example: 31% of posts take the form of the question leading up to a mean joke. There is still need for a representative and syntactically diverse implicit hate benchmark.

Our contribution is similar to the Gab Hate Corpus of Kennedy et al. (2018), which provides both explicit and implicit hate and target labels for a random sample of 27K Gab messages. We extend this work with a *theoretically-grounded taxonomy* and *fine-grained labels for implicit hate speech* beyond the umbrella categories, Assault on Human Dignity (HD) and Call for Violence (CV). Following the work of Sap et al. (2020), we provide free-text annotations to capture messages' pragmatic implications. However, we are the first to take this framework, which was originally applied stereotype bias, and extend it to *implicit* hate speech more broadly. Implicitly stereotypical language is just a subset of the implicit hate we cover, since we also include other forms of sarcasm, intimidation or incitement to violence, hidden threats, white grievance, and subtle forms of misinformation. Our work also complements recent efforts to capture and understand *microaggressions* (Breitfeller et al., 2019), a similarly elusive class that draws on subtle and *unconscious* linguistic reflections of social bias, prejudice and inequality (Sue, 2010). Similar to Breitfeller et al. (2019), we provide a representative and domain-general typology and dataset, but ours are more representative of *active hate groups* in the United States, and our definitions extend to *intentionally* veiled acts of intimidation, threats, and abuse.

3 Taxonomy of Implicit Hate Speech

Implicit hate speech is a subclass of hate speech defined by *the use of coded or indirect language* such as sarcasm, metaphor and circumlocution to disparage a protected group or individual, or to convey prejudicial and harmful views about them (Gao et al., 2017; Waseem et al., 2017). The NLP community has not yet confronted, in a consistent and unified manner, the multiplicity of subtle challenges that implicit hate presents for online communities. To this end, we introduce a new typology

[1] We use the swear word list from https://bit.ly/2SQySZv, excluding ambiguous terms like *bloody*, *prick*, etc.

Work	Source	Domain / Scope	Size	Balance	Expletives	Public	Target	Implicit	Implied
Basile et al. (2019)	Twitter	Misogynistic, anti-immigrant	19,600	Unknown	Unknown	✓	✓		
Burnap and Williams (2014)	Twitter	Lee Rigby murder	1,901	11.7%	Unknown				
Davidson et al. (2017)	Twitter	HateBase terms	24,802	5.0%	69.8%	✓			
Djuric et al. (2015)	Yahoo Finance	Unknown	951,736	5.9%	Unknown				
Founta et al. (2018)	Twitter	Offensive terms	80,000	7.5%	73.9%	✓			
Gao and Huang (2017)	Fox News Comments	Unknown	1,528	28.5%	Unknown				
de Gibert et al. (2018)	Stormfront	One hate group	9,916	11.3%	7.8%	✓			
Kennedy et al. (2018)	Gab	Random sample	27,665	9.1%	28.2%	✓	✓	✓	
Sap et al. (2020)	Compilation	Mixed	44,671	44.8%	28.5%	✓	✓		✓
Warner and Hirschberg (2012)	Yahoo + Web	Anti-semitic	9,000	Unknown	Unknown	✓			
Waseem and Hovy (2016)	Twitter	Sexist, racist terms	16,914	31.7%	17.6%	✓			
Zampieri et al. (2019)	Twitter	Political phrases	14,000	32.9%	Unknown	✓	✓		
Implicit Hate Corpus (ours)	Twitter	Hate groups	22,584	39.6%	3.2%	✓	✓	✓	✓

Table 1: Summary of English hate speech datasets in terms of *Domain / Scope*, *Size*, hate class *Balance* ratio, the proportion of *Expletives* in the hate class, and the inclusion of *Target* demographic, binary *Implicit* hate speech labels, and *Implied* statement summaries. Most datasets cover a narrow subset of hate speech like anti-semitism or sexism, and do not include implicit hate labels. Ours is the first to include a fine-grained implicit hate taxonomy.

for characterizing and detecting different forms of implicit hate, based on social science and relevant NLP literature. Our categories are not necessarily mutually exclusive, but they represent principle axes of implicit hate, and while they may not be collectively exhaustive, we find they cover 98.6% of implicit hate in a representative sample of the most prevalent hate ideologies in the U.S.

White Grievance includes frustration over a minority group’s perceived privilege and casting majority groups as the real victims of racism (Berbrier, 2000; Bloch et al., 2020). This language is linked to extremist behavior and support for violence (Miller-Idriss, 2020). An example is *Black lives matter and white lives don’t? Sounds racist.*

Incitement to Violence includes flaunting in-group unity and power or elevating known hate groups and ideologies (Somerville, 2011). Phrases like *‘white brotherhood* operate in the former manner, while statements like *Hitler was Germany – Germans shall rise again!* operate in the latter, elevating nationalism and Nazism. Article 20 of the UN International Covenant on Civil and Political Rights (Assembly, 1966) states that speech which incites violence shall be prohibited by law.

Inferiority Language implies one group or individual is inferior to another (Nielsen, 2002), and it can include dehumanization (denial of a person’s humanity), and toxification (language that compares the target with disease, insects, animals), both of which are early warning signs of genocide (Leader Maynard and Benesch, 2016; Neilsen, 2015). Inferiority language is also related to *assaults on human dignity* (Kennedy et al., 2018), *dominance* (Saha et al., 2018), and *declarations of superiority of the in-group* (Fortuna and Nunes, 2018). For example, *It’s not a coincidence the best places to live are majority white.*

Irony refers to the use of sarcasm (Waseem and Hovy, 2016; Justo et al., 2014), humor (Fortuna and Nunes, 2018), and satire (Sanguinetti et al., 2018) to attack or demean a protected class or individual. For example, in the context of one hate group, the tweet *Horrors... Disney will be forced into hiring Americans* works to discredit Disney for allegedly hiring only non-citizens or, really, non-whites. Irony is not exempt from our hate speech typology, since it is commonly used by modern online hate groups to mask their hatred and extremism (Dreisbach, 2021).

Stereotypes and Misinformation associate a protected class with negative attributes such as crime or terrorism (Warner and Hirschberg, 2012; Sanguinetti et al., 2018) as in the rhetorical question, *Can someone tell the black people in Chicago to stop killing one another before it becomes Detroit?* This class also includes misinformation that feeds stereotypes and vice versa, like holocaust denial and other forms of historical negationism (Belavusau, 2017; Cohen-Almagor, 2009).

Threatening and Intimidation convey a speaker commitment to a target's pain, injury, damage, loss, or violation of rights. While explicitly violent threats are well-recognized in the hate speech literature (Sanguinetti et al., 2018), here we highlight threats related to implicit violation of rights and freedoms, removal of opportunities, and more subtle forms of intimidation, such as *All immigration of non-whites should be ended.*

4 Data Collection and Annotation

We collect and annotate a benchmark dataset for implicit hate language using our taxonomy. Our main source of data uses content published by online hate groups and their followers on Twitter for two reasons. First, as modern hate groups have become more active online, they provide an increasingly vivid picture of the more subtle and coded forms of hate that we are interested in. Second, the problem of hateful misinformation is compounded on social media platforms like Twitter where around 3 out of 4 users get their news (Shearer and Gottfried, 2017). This motivates a representative sample of *online* communication exchanged on *Twitter* between members of the *most prominent U.S hate groups*.

We focus on the eight largest ideological clusters of U.S. hate groups as given by the SPLC (2019) report. These ideological classes are *Black Separatist* (27.1%), *White Nationalist* (16.4%), *Neo-Nazi* (6.2%), *Anti-Muslim* (8.9%), *Racist Skinhead* (5.1%), *Ku Klux Klan* (5.0%), *Anti-LGBT* (7.4%), and *Anti-Immigrant* (2.12%). Detailed background and discussion on each hate ideology can be found at the the SPLC Extremist Files page (SPLC, 2020).

4.1 Data Collection and Filtering

We matched all SPLC hate groups with their corresponding Twitter accounts using the account names and bios. Then, for each ideological cluster above, we selected the three hate group accounts with the most followers, since these were likely to be the most visible and engaged. We collected all tweets, retweets, and replies from the timelines of our selected hate groups between January 1, 2015 and December 31, 2017, for a total of 4,748,226 tweets, giving us with an broad sample of hate group activity before many accounts were banned.

Hateful content is semantically diverse, with different hate groups motivated by different ideologies. Seeking a representative sample, we identified group-specific salient content from each ideology by performing part of speech (POS) tagging on each tweet. Then we computed the log odds ratio with informative Dirichlet prior (Monroe et al., 2008) for each noun, hashtag, and adjective to identify the top 25 words per ideology. After filtering for tweets that contained one of the salient keywords, we ran the 3-way HateSonar classifier of Davidson et al. (2017) to remove content that was likely to be explicitly hateful. Specifically, we removed all tweets that were classified as *offensive*, and then ran a final sweep over the *neutral* and *hate* categories, removing tweets that contained any explicit keyword found in NoSwear (Jones, 2020) or Hatebase (Hatebase, 2020).

4.2 Crowdsourcing and Expert Annotation

To acquire implicit hate speech labels with two different resolutions, we ran two stages of annotation. First, we collected high-level labels, *explicit hate, implicit hate,* or *not hate*. Then, we took a second pass through the implicit hate tweets with expert annotation over the fine-grained implicit hate taxonomy from Section 3.

4.2.1 Stage 1: High Level Categorization

Amazon Mechanical Turk (MTurk) annotators completed our high-level labeling task. We provided them with a definition of hate speech (Twitter, 2021) and examples of explicit, implicit, and non-hateful content (See Appendix A), and required them to pass a short five-question qualification check for understanding with a score of at least 90% in accordance with crowdsourcing standards (Sheehan, 2018). We paid annotators a fair wage above the federal minimum. Three workers labeled each tweet, and they reached majority agreement for 95.3% of tweets, with perfect agreement on 45.6% of the data. The Intraclass Correlation for one-way random effects between $k = 118$ raters was $ICC(1, k) = 0.616$, which indicates moderate inter-rater agreement. Using the majority vote, we obtained consensus labels for 19,112 labeled tweets in total: 933 *explicit hate*, 4,909 *implicit hate*, and 13,291 *not hateful* tweets.

4.2.2 Stage 2: Fine-Grained Implicit Hate

To promote a more nuanced understanding of our 4,909 implicit hate tweets, we labeled them using our fine-grained category definitions in Section 3, adding *other* and *not hate* to take care of any other situations. Since these fine-grained categories were too subtle for MTurk workers,[2] we hired three research assistants to be our expert annotators. We trained them over multiple sessions by walking them through seven small pilot batches and resolving disagreements after each test until they reached moderate agreement. On the next round of 150 tweets, their independent annotations reached a Fleiss' Kappa of 0.61. Each annotator then continued labeling an independent partition of the data. Halfway through this process, we ran another attention check with 150 tweets and found that agreement remained consistent with a Fleiss' Kappa of 0.55. Finally, after filtering out tweets marked as *not hate*, there were 4,153 labeled implicit hate tweets remaining. The per-category statistics are summarized in the *# Tweets Pre Expn.* column of Table 2.

4.2.3 Corpus Expansion

Extreme class imbalance may challenge implicit hate classifiers. To address this disparity, we expand the minority classes, both with bootstrapping and out-of-domain samples.

For bootstrapping, we trained a 6-way BERT classifier on the 4,153 implicit hate labels in the manner of Section 5.1 and ran it on 364,300 unlabeled tweets from our corpus. Then we randomly sampled 1,800 tweets for each of the three minority classes according to the classifications *inferiority*, *irony*, and *threatening*. Finally, we augmented this expansion with out-of-domain (OOD) samples from Kennedy et al. (2018) and Sap et al. (2020). By drawing both from OOD and bootstrapped in-domain samples, we sought to balance two key limitations: (1) bootstrapped samples may be inherently easier, while (2) OOD samples contain artifacts that allow models to benefit from spurious correlations. Our expert annotators labeled this data, and by adding the minority labels from this process, we improved the class balance for a total of 6,346 implicit tweets shown in the *# Tweets Post Expn.* column of Table 2.

4.2.4 Hate Targets and Implied Statement

For each of the 6,346 implicit hate tweets, two separate annotators provided us with the message's *target demographic group* and its *implied statement* in free-text format. Implied statements were formatted as Hearst-like patterns (Indurkhya and Damerau, 2010) of the form <*target*> *{do, are, commit}* <*predicate*>, where <*target*> might be phrases such as *immigrants, black folks*.

Label	**# Tweets Pre Expn**	**# Tweets Post Expn**	**% Post Expn**
Grievance	1,455	1,538	24.2%
Incitement	1,176	1,269	20.0%
Inferiority	241	863	13.6%
Irony	134	797	12.6%
Stereotypical	1,032	1,133	17.9%
Threatening	57	666	10.5%
Other	58	80	1.2%
Total	4,153	6,346	100%

Table 2: Implicit hate category label distribution before and after the expansion stage

5 Implicit Hate Speech Classification

We experiment with two classification tasks: (1) distinguishing implicit hate speech from non-hate, and (2) categorizing implicit hate speech using one of the 6 classes in our fine-grained taxonomy.

5.1 Experimental Setup

Using a 60-20-20 split for each task, we trained, validated, and tested SVM and BERT baselines. We tried standard unigrams, TF-IDF, and Glove embedding (Pennington et al., 2014) features and tuned linear SVMs with $C \in \{0.1, 1, 10, 100, 1000\}$. Next, we fine-tuned BERT with the learning rate in {2e-5, 3e-5, 5e-5} and the number of epochs in $\{1, 2, 3, 4\}$.[3] We also balanced the training data (**BERT + Aug**) with back-translation from Russian via FairSeq (Gehring et al., 2017), using a grid search over the sampling temperature in {0.5, 0.7, 0.9}. Finally, we supplemented the previous methods with knowledge-based features to learn implicit associations between entities. In detail, we matched tweets to entities like *white people*, *Islam*, and *antifa* from Wikidata Knowledge Graph (Vrandečić and Krötzsch, 2014) (**BERT + Aug + Wikidata**) and ConceptNet numberbatch (Speer et al., 2017) (**BERT + Aug + ConceptNet**) by string-matching unigrams, bigrams, and trigrams. Then we averaged across the pre-trained entity embeddings matched for each message.[4] Finally,

[2] We saw less than 30% agreement when we ran this task over three batches of around 200 tweets each on MTurk.

[3] We kept $\epsilon = 1.0 \times 10^{-8}$ and the batch size fixed at 8

[4] 11,163 / 22,584 tweets ($\approx 54\%$) were matched to one Wikidata entity (none were matched to more than one); 22,554 / 22,584 tweets ($> 99\%$) were matched to at least one ConceptNet entity, and the average number of matches per tweet

	Binary Classification				Implicit Hate Categories			
Models	P	R	F	Acc	P	R	F	Acc
Hate Sonar	39.9	48.6	43.8	54.6	-	-	-	-
Perspective API	50.1	61.3	55.2	63.7	-	-	-	-
SVM (n-grams)	61.4	67.7	64.4	72.7	48.8	49.2	48.4	54.2
SVM (TF-IDF)	59.5	68.8	63.9	71.6	53.0	51.7	51.5	56.5
SVM (GloVe)	56.5	65.3	60.6	69.0	46.8	48.9	46.3	51.3
BERT	**72.1**	66.0	68.9	**78.3**	**59.1**	57.9	58.0	62.9
BERT + Aug	67.8	**73.2**	**70.4**	77.5	58.6	**59.1**	**58.6**	**63.8**
BERT + Aug + Wikidata	67.6	72.3	69.9	77.3	53.9	55.3	54.4	62.8
BERT + Aug + ConceptNet	68.6	70.0	69.3	77.4	54.0	55.4	54.3	62.5

Table 3: Classification performance metrics averaged over five random seeds. (*Left*) **Binary Classification**. Performance metrics for implicit hate vs. not hate classification. (*Right*) **Implicit Hate Categories**. Macro performance metrics for *fine-grained category* classification via implicit hate taxonomy. Best performance is bolded.

we concatenated the 768-dimensional BERT final layer with the 200-dimensional Wikidata (or 300-dimensional ConceptNet) embeddings, and fed this representation into an MLP with two hidden layers of dimension 100 and ReLU activation between them, using categorical Cross Entropy loss.

5.2 Implicit Hate Classification Results

In binary implicit hate speech classification on the left side of Table 3, baseline SVM models offer competitive performance with F_1 scores up to 64.4, while the fine-tuned neural models gain up to 6 additional points. The BERT-base model achieves significantly better macro precision than the linear SVMs (72.1 vs. at most 61.4), demonstrating a compositional understanding beyond simple keyword-matching. When we look at our best **BERT + Aug** model, the implicit category most confused with non-hate was *Incitement* (36.3% of testing examples were classified as not hate), followed by *White Grievance* (29.6%), *Stereotypical* (23.3%), *Inferiority* (12.3%), *Irony* (9.3%), and *Threatening* (5.5%). In our 6-way classification task on the right of Table 3, we find that the BERT-base models again outperform the linear models. Augmentation does not significantly improve performance in either task since our data is already well-balanced and representative. Interestingly, integrating Wikidata and ConceptNet did not lead to any performance boost either. This suggests detecting implicit hate speech might require more compositional reasoning over the involved entities and we urge future work to investigate this. For additional comparisons, we consider a zero-shot setting where we test Google's Perspective API[5] and the HateSonar classifier of Davidson et al. (2017). Our fine-tuned baselines significantly outperform both zero-shot baselines, which were trained on explicit hate.

5.3 Challenges in Detecting Implicit Hate

To further understand the challenges of implicit hate detection and promising directions for future work, we investigated 100 randomly sampled false negative errors from our best model in the binary task (BERT+Aug) and found a set of linguistic classes it struggles with.[6] **(1) Coded hate symbols** (Qian et al., 2019) such as *#WPWW* (white pride world wide), *#NationalSocialism* (Nazism), and *(((they)))* (an anti-Semitic symbol) are contained in 15% of instances, and our models fail to grasp their semantics. While individual sentences appear harmless, implicit hate can occur in **(2) discourse relations** (de Gibert et al., 2018) (19% of instances) like the implied causal relation between the conjunction *I like him* and *he's white*. Additionally, misinformation (Islam et al., 2020) and out-group **(3) entity framing** (Phadke and Mitra, 2020) (25%) can be context-sensitive, as in the headline *three Muslims convicted*. Even positive framing of a negative entity can be problematic, like describing a Nazi soldier as *super cool*.

Inferiority statements like *POC need us and not the other way around* also require a deep under-

was 14.

[5] https://www.perspectiveapi.com/

[6] For robustness check, we also labeled 100 false positives from the BERT base model and found the distribution of errors remains similar.

	Target Group				Implied Statement			
Models	BLEU	BLEU*	Rouge-L	Rouge-L*	BLEU	BLEU*	Rouge-L	Rouge-L*
GPT-gdy	43.7	65.2	42.9	63.3	41.1	58.2	31	45.3
GPT-top-p	57.7	76.8	55.8	74.6	55.2	69.4	40	53.9
GPT-beam	59.3	81	57.3	78.6	57.8	73.8	46.5	63.4
GPT-2-gdy	45.3	67.6	44.6	66	42.3	59.3	32.7	47.4
GPT-2-top-p	58.0	76.9	56.2	74.8	55.1	69.3	39.6	53.1
GPT-2-beam	**61.3**	**83.9**	**59.6**	**81.8**	**58.9**	**75.3**	**48.3**	**65.9**

Table 4: Evaluation of the generation models for Target Group and Implied Statement. (*) denotes the maximum versus the average score (without asterisk). gdy: greedy decoding, beam: beam search with 3 hypotheses, and top-p: nucleus sampling with $p = 0.92$

standing of **(4) commonsense** (11%) surrounding social norms (e.g. *a dependant is inferior to a supplier*) (Forbes et al., 2020). Other challenge cases contain highly **(5) metaphorical language** (7%), like the animal metaphor in *a world without white people : a visual look at a mongrel future*. **(6) Colloquial or idiomatic speech** (17%) appears in subtle phrases like *infrastructure is the white man's game*, and **(7) Irony** (15%) detection (Waseem and Hovy, 2016) may require pragmatic reasoning and understanding, such as in the phrase *hey kids, wanna replace white people.*

When we sample false positives, we find our models are prone to **(8) identity term bias** (Dixon et al., 2018). Given the high density of identity terms like *Jew* and *Black* in hateful contexts, our models overclassified tweets with these terms as *hateful*, and particularly *stereotypical* speech. In a similar manner, our model also incorrectly associated *white grievance* with all *diversity*-related discourse, *incitement* with controversial topics like *war* and *race*, and *inferiority language* with value-laden terms like *valid* and *wealth.*

To sum up, our dataset contains rich linguistic phenomena and an array of subtleties that challenge current state-of-the-art baselines, which can serve as a useful benchmark and offer multiple new directions for future work.

6 Explaining Implicit Hate Speech

This section presents our generation results for natural language explanations of both (1) who is being targeted and (2) what the implied message is for each implicitly hateful tweet. Generating such explanations can help content moderators better understand the severity and nature of automatically-flagged messages. Additionally, we echo efforts from social media companies (e.g., Instagram (Bryant, 2019)) where the application alerts the user when the post is flagged "offensive," and asks them if they really want to post it. This strategy has proven successful in deterring hurtful comments. Our work could inspire a similar strategy for implicit hate speech. By showing the user the implied meaning of their post before it is posted, we would enable them to recognize the severity of their words and possibly reconsider their decision to post.

6.1 Task Formulation

Our goal is to develop a natural language system that, given a post, generates a hateful post's intended target and hidden implied meanings. Therefore, we formulate the problem as a conditional generation task (i.e., conditioned on the post content). During training, the generation model takes a sequence of tokens as input:

$$\begin{aligned}\mathbf{x} =&\{[STR], t_1, t_2, ..., t_n, [SEP],\\ &t_{[G1]}, t_{[G2]}, ..., [SEP], t_{[S1]}, t_{[S2]}, ..., [END]\}\end{aligned}$$

with start token [STR], tweet tokens $t_1 : t_n$, target group $t_{[Gi]}$, and implied statement $t_{[Si]}$, and minimizes the cross-entropy loss $-\sum_l \log P(\tilde{t}_l | t_{<l})$.

During inference, our goal is to mimic real-world scenarios when only the post is available. Therefore, the input to the model only contains post tokens $t_1 : t_n$ and we experiment with multiple decoding strategies: greedy search (gdy), beam search, and top-p (nucleus) sampling to generate the explanations $t_{[G_i]}$ and $t_{[S_i]}$.

6.2 Experiment Setup

Our ground-truth comes from the free-text *target demographic* and *implied statement* annotations that we collected for all 6,346 implicit hate tweets

Post	Type	Target Group	Implied Statement
(a) white revolution is the only solution	incitement	non-white people *white people*	non-white people are inferior *white people are superior to other races*
(b) yes you are fine in a white majority country. how is mexico doing these days? people come here because you couldn't build it.	inferiority	mexican people *mexicans*	mexican people do not build things *mexicans are incompetent*
(c) black people only seem to thrive in white countries. how strange	irony	black people *black people*	black people are inferior *black people can't survive without white people*
(d) soon white people will not be a majority anywhere. you believe	grievance	whites *minorities*	whites have been driven out of the country *minorities want to take over the world*
(e) muslims are always wanting to kill someone! its their duty to kill the infidel.	stereotypical	muslims *muslims*	muslims are killers *muslims are violent*
(f) oh how i wish a man with a gun was there! wake up white people! they hate us	threatening	white people *non-whites*	white people are hated *non-whites hate whites*

Table 5: Example posts from our dataset along with their implicit category labels, the GPT-2 generated target and implied statements (first row of each block), and the ground truth target and implied statements (final row of each block, in *italics*). Generated implied statements are semantically similar to the ground truth statements.

in Section 4.2.4, with 75% for training, 12.5% for validation, and 12.5% for testing. Since we collect multiple annotations for each post (2 per tweet), we ensure that each post and its corresponding annotations belongs only to one split.

Following recent work on social bias inference and commonsense reasoning (Sap et al., 2020; Forbes et al., 2020; Sharma et al., 2020; Bosselut et al., 2019), we fine-tune Open-AI's GPT (Radford et al., 2018) and GPT-2 (Radford et al., 2019) pre-trained language models to the task and evaluate using BLEU (Papineni et al., 2002) and ROUGE-L (Lin, 2004).

We pick BLEU since it is standard for evaluating machine translation models and ROUGE which is used in summarization contexts; both have been adopted extensively in prior literature. These automatic metrics indicate the quality of the generated target group and implied statement compared to our annotated ground-truth in terms of n-grams and the longest common sequence overlaps. Since there are two ground truth annotations per tweet, we measure both the averaged metrics across both references, and the maximum metrics (BLEU* and ROUGE-L*).

We tuned hyperparameters and selected the best models based on their performance on the development set, and we reported evaluation results on the test.[7] For decoding, we generate one frame for greedy decoding and three hypotheses for beam search and top-p (nucleus) sampling with $p = 0.92$ and choose the highest scoring frame.

6.3 Generation Results

In Table 4 we find that, GPT-2 outperforms GPT in both *target group* and *implied statement* generation. This difference is likely because GPT-2 was trained on English web text while GPT was trained on fiction books and web text is more similar to our domain. The BLEU and ROUGE-L scores are higher for the *target group* (e.g., 83.9 BLEU) than for the *implied statement* (e.g., 75.3 BLEU), consistently across both averaged and maximum scores. This is likely because the implied statement is longer, more nuanced, and less likely to be contained in the text itself. Additionally, beam search achieves the highest performance for both GPT and GPT-2, followed by top-p. This is not surprising since both decoding strategies consider multiple hypotheses. Since BLEU and ROUGE-L measure word overlap and not semantics, it is possible that the results in Table 4 are overly pessimistic. The GPT-2 generated implied statements in Table 4 actually describe the complement (a,d), generalization (b), extrapolation (c), or paraphrase (e,f) of the ground truth, and are thus aligned, despite differences in word choice. Overall, our generation results are promising. Transformer-based models may play a key role in explaining the severity and nature of online implicit hate.

7 Conclusion

In this work, we introduce a theoretical taxonomy of implicit hate speech and a large-scale benchmark corpus with fine-grained labels for each message and its implication. As an initial effort, our work enables the NLP communities to better understand and model implicit hate speech at scale. We also provide several state-of-the-art baselines for detect-

[7] We fine-tune for $e \in \{1, 2, 3, 5\}$ epochs with a batch size of 2 and learning rate of 5×10^{-5} with linear warm up

ing and explaining implicit hate speech. Experimental results show these neural models can effectively categorize hate speech and spell out more fine-grained implicit hate speech and explaining these hateful messages.

Additionally, we identified eight challenges in implicit hate speech detection: coded hate symbols, discourse relations, entity framing, commonsense, metaphorical language, colloquial speech, irony, and identity term bias. To mitigate these challenges, future work could explore deciphering models for coded language (Kambhatla et al., 2018; Qian et al., 2019), lifelong learning of hateful language (Qian et al., 2021), contextualized sarcasm detection, and bias mitigation for named entities in hate speech detection systems (Xia et al., 2020) and their connection with our dataset.

We demonstrate that our corpus can serve as a useful research benchmark for understanding implicit hate speech online. Our work also has implications towards the emerging directions of countering online hate speech (Citron and Norton, 2011; Mathew et al., 2019), detecting online radicalization (Ferrara et al., 2016) and modeling societal systematic racism, prejudicial expressions, and biases (Davidson et al., 2019; Manzini et al., 2019; Blodgett et al., 2020).

Ethical Considerations

This study has been approved by the Institutional Review Board (IRB) at the researchers' institution. For the annotation process, we included a warning in the instructions that the content might be offensive or upsetting. Annotators were also encouraged to stop the labeling process if they were overwhelmed. We also acknowledge the risk associated with releasing an implicit hate dataset. However, we believe that the benefit of shedding light on the implicit hate phenomenon outweighs any risks associated with the dataset release.

Acknowledgments

The authors would like to thank the members of SALT lab and the anonymous reviewers for their thoughtful feedback. The annotation process was funded through the School of Interactive Computing at Georgia Tech and the Institute for Data Engineering and Science (IDEAS) Data Curation Award to ElSherief and Yang. The work is supported in part by Russell Sage Foundation.

References

UN General Assembly. 1966. International covenant on civil and political rights. *United Nations, Treaty Series*, 999:171.

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 759–760.

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

Uladzislau Belavusau. 2017. Hate speech. *Max Planck Encyclopedia of Comparative Constitutional Law (Oxford University Press, 2017 Forthcoming)*.

Mitch Berbrier. 2000. The victim ideology of white supremacists and white separatists in the united states. *Sociological Focus*, 33(2):175–191.

Katrina Rebecca Bloch, Tiffany Taylor, and Karen Martinez. 2020. Playing the race card: White injury, white victimhood and the paradox of colour-blind ideology in anti-immigrant discourse. *Ethnic and Racial Studies*, 43(7):1130–1148.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy. Association for Computational Linguistics.

Luke Breitfeller, Emily Ahn, David Jurgens, and Yulia Tsvetkov. 2019. Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1664–1674, Hong Kong, China. Association for Computational Linguistics.

Miranda Bryant. 2019. Instagram's anti-bullying ai asks users: 'are you sure you want to post this?'. *The Guardian*.

Peter Burnap and Matthew Leighton Williams. 2014. Hate speech, machine classification and statistical modelling of information flows on twitter: Interpretation and communication for policy decision making. *Pre-print*.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, Inga Kartoziya, and Michael Granitzer. 2020. I feel offended, don't be abusive! implicit/explicit messages in offensive and abusive language. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6193–6202, Marseille, France. European Language Resources Association.

Danielle Keats Citron and Helen Norton. 2011. Intermediaries and hate speech: Fostering digital citizenship for our information age. *Boston University Law Review*, 91:1435.

Raphael Cohen-Almagor. 2009. Holocaust denial is a form of hate speech. In *Amsterdam Law Forum*, volume 2, pages 33–42.

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

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *ArXiv preprint*, abs/1703.04009.

Quentin Dénigot and Heather Burnett. 2020. Dogwhistles as identity-based interpretative variation. In *Proceedings of the Probability and Meaning Conference (PaM 2020)*, pages 17–25, Gothenburg. Association for Computational Linguistics.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73.

Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. 2015. Hate speech detection with comment embeddings. In *Proceedings of the 24th international conference on world wide web*, pages 29–30.

Tom Dreisbach. 2021. How extremists weaponize irony to spread hate.

Emilio Ferrara, Wen-Qiang Wang, Onur Varol, Alessandro Flammini, and Aram Galstyan. 2016. Predicting online extremism, content adopters, and interaction reciprocity. In *International conference on social informatics*, pages 22–39. Springer.

Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. Social chemistry 101: Learning to reason about social and moral norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 653–670, Online. Association for Computational Linguistics.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30.

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

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

Lei Gao, Alexis Kuppersmith, and Ruihong Huang. 2017. Recognizing explicit and implicit hate speech using a weakly supervised two-path bootstrapping approach. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 774–782, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1243–1252. PMLR.

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

Joshua R Gubler and Nathan P Kalmoe. 2015. Violent rhetoric in protracted group conflicts: Experimental evidence from israel and india. *Political Research Quarterly*, 68(4):651–664.

Hatebase. 2020. [link].

Nitin Indurkhya and Fred J Damerau. 2010. *Handbook of natural language processing*, volume 2. CRC Press.

Md Rafiqul Islam, Shaowu Liu, Xianzhi Wang, and Guandong Xu. 2020. Deep learning for misinformation detection on online social networks: a survey and new perspectives. *Social Network Analysis and Mining*, 10(1):1–20.

Ryan Jones. 2020. List of swear words, bad words, & curse words - starting with a.

David Jurgens, Libby Hemphill, and Eshwar Chandrasekharan. 2019. A just and comprehensive strategy for using NLP to address online abuse. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3658–3666, Florence, Italy. Association for Computational Linguistics.

Raquel Justo, Thomas Corcoran, Stephanie M Lukin, Marilyn Walker, and M Inés Torres. 2014. Extracting relevant knowledge for the detection of sarcasm and nastiness in the social web. *Knowledge-Based Systems*, 69:124–133.

Nishant Kambhatla, Anahita Mansouri Bigvand, and Anoop Sarkar. 2018. Decipherment of substitution ciphers with neural language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 869–874, Brussels, Belgium. Association for Computational Linguistics.

Brendan Kennedy, Mohammad Atari, Aida M Davani, Leigh Yeh, Ali Omrani, Yehsong Kim, Kris Coombs Jr, Shreya Havaldar, Gwenyth Portillo-Wightman, Elaine Gonzalez, et al. 2018. The gab hate corpus: A collection of 27k posts annotated for hate speech. *PsyArXiv. July*, 18.

David Klepper. 2020. Facebook removes nearly 200 accounts tied to hate groups.

Jonathan Leader Maynard and Susan Benesch. 2016. Dangerous speech and dangerous ideology: An integrated model for monitoring and prevention. *Genocide Studies and Prevention*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Rijul Magu and Jiebo Luo. 2018. Determining code words in euphemistic hate speech using word embedding networks. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 93–100, Brussels, Belgium. Association for Computational Linguistics.

Thomas Manzini, Lim Yao Chong, Alan W Black, and Yulia Tsvetkov. 2019. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. 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)*, pages 615–621, Minneapolis, Minnesota. Association for Computational Linguistics.

Binny Mathew, Punyajoy Saha, Hardik Tharad, Subham Rajgaria, Prajwal Singhania, Suman Kalyan Maity, Pawan Goyal, and Animesh Mukherjee. 2019. Thou shalt not hate: Countering online hate speech. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 369–380.

Cynthia Miller-Idriss. 2020. *Hate in the homeland: The new global far right*. Princeton University Press.

Burt L Monroe, Michael P Colaresi, and Kevin M Quinn. 2008. Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403.

Rhiannon S Neilsen. 2015. 'toxification'as a more precise early warning sign for genocide than dehumanization? an emerging research agenda. *Genocide Studies and Prevention: An International Journal*, 9(1):9.

Laura Beth Nielsen. 2002. Subtle, pervasive, harmful: Racist and sexist remarks in public as hate speech. *Journal of Social Issues*, 58(2):265–280.

John T Nockleby. 2000. Hate speech. *Encyclopedia of the American constitution*, 3(2):1277–1279.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Shruti Phadke and Tanushree Mitra. 2020. Many faced hate: A cross platform study of content framing and information sharing by online hate groups. In *CHI '20: CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, April 25-30, 2020*, pages 1–13. ACM.

James A Piazza. 2020. Politician hate speech and domestic terrorism. *International Interactions*, pages 1–23.

Jing Qian, Mai ElSherief, Elizabeth Belding, and William Yang Wang. 2019. Learning to decipher hate symbols. 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)*, pages 3006–3015, Minneapolis, Minnesota. Association for Computational Linguistics.

Jing Qian, Hong Wang, Mai ElSherief, and Xifeng Yan. 2021. Lifelong learning of hate speech classification on social media. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human*

Language Technologies, pages 2304–2314, Online. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding with unsupervised learning. *Technical report, OpenAI*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Punyajoy Saha, Binny Mathew, Pawan Goyal, and Animesh Mukherjee. 2018. Hateminers: detecting hate speech against women. *ArXiv preprint*, abs/1812.06700.

Manuela Sanguinetti, Fabio Poletto, Cristina Bosco, Viviana Patti, and Marco Stranisci. 2018. An italian twitter corpus of hate speech against immigrants. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

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

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

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

Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5263–5276, Online. Association for Computational Linguistics.

Elisa Shearer and Jeffrey Gottfried. 2017. News use across social media platforms 2017. *Pew Research Center*, 7(9).

Kim Bartel Sheehan. 2018. Crowdsourcing research: data collection with amazon's mechanical turk. *Communication Monographs*, 85(1):140–156.

Leandro Silva, Mainack Mondal, Denzil Correa, Fabrício Benevenuto, and Ingmar Weber. 2016. Analyzing the targets of hate in online social media. In *10th International AAAI Conference on Web and Social Media*, pages 687–690. AAAI.

Keith Somerville. 2011. Violence, hate speech and inflammatory broadcasting in kenya: The problems of definition and identification. *Ecquid Novi: African Journalism Studies*, 32(1):82–101.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 4444–4451. AAAI Press.

SPLC. 2019. Hate map.

SPLC. 2020. Ideologies.

Derald Wing Sue. 2010. *Microaggressions in everyday life: Race, gender, and sexual orientation*. John Wiley & Sons.

Twitter. 2021. Twitter's policy on hateful conduct | twitter help.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

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

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

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

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 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)*, pages 602–608, Minneapolis, Minnesota. Association for Computational Linguistics.

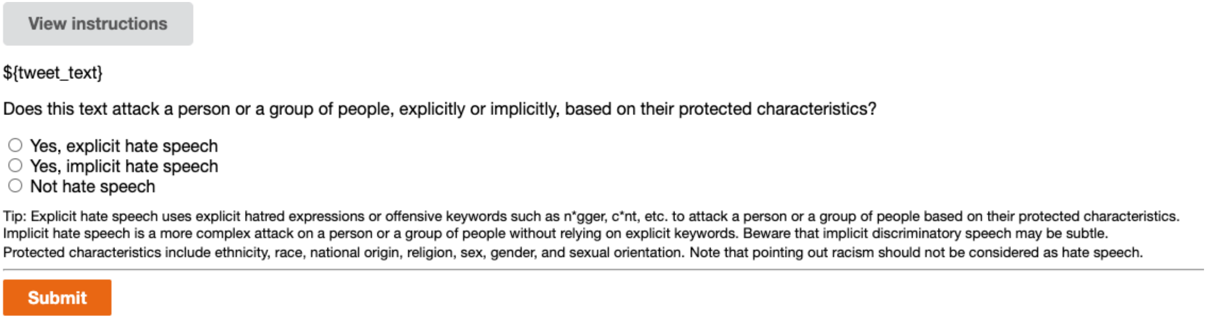
Mengzhou Xia, Anjalie Field, and Yulia Tsvetkov. 2020. Demoting racial bias in hate speech detection. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, pages 7–14, Online. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. 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)*, pages 1415–1420, Minneapolis, Minnesota. Association for Computational Linguistics.

A Data Collection Details

In our first annotation stage (Section 4.2.1), we provide a broad definition of hate speech grounded in Twitter's hateful conduct policy (Twitter, 2021), and detailed definitions for what constitutes explicit hate, implicit hate, and non-hateful content with examples from each class. We explain that explicit hate speech contains explicit keywords directed towards a protected entity. We define implicit hate speech as outlined in the paper and ground this definition in a quote from Lee Atwater on how discourse can appeal to racists without sounding racist: "*You start out in 1954 by saying, "N*gger, n*gger, n*gger." By 1968 you can't say "n*gger"—that hurts you, backfires. So you say stuff like, uh, forced busing, states' rights, and all that stuff, and you're getting so abstract*". To ensure quality, we chose only AMT Master workers who (1) have approval rate >98% and more than 5000 HITs approved, (2) scored $\geq$ 90% on our five-question qualification test where they must (a) identify the differences between explicit and implicit hate speech and (b) identify the hate target even if the target is not explicitly mentioned. Figures 2 and 4 depict snippets of the first stage annotation task and the instructions provided to guide the annotators, respectively.

For the second-stage annotation (Section 4.2.2), we observed the following per-category kappa scores at the beginning/middle: (threatening, 1.00/0.66), (stereotypical, 0.67/0.55), (grievance, 0.61/0.63), (incitement, 0.63/0.53), (not hate, 0.55/0.54), (inferiority, 0.47/0.41), and (irony, 0.40/0.31). Even in the worst case, there was fair to moderate agreement. We will add these metrics to the Appendix. The total annotation cost for Stage 1 and 2 was $15k. Limited by our budget, we chose to employ expert annotators to label independent portions of the data once we observed fair to substantial agreement among them. Figure 3 depicts a snippet of the hate target and implied statement data collection for each implicit hate speech post.



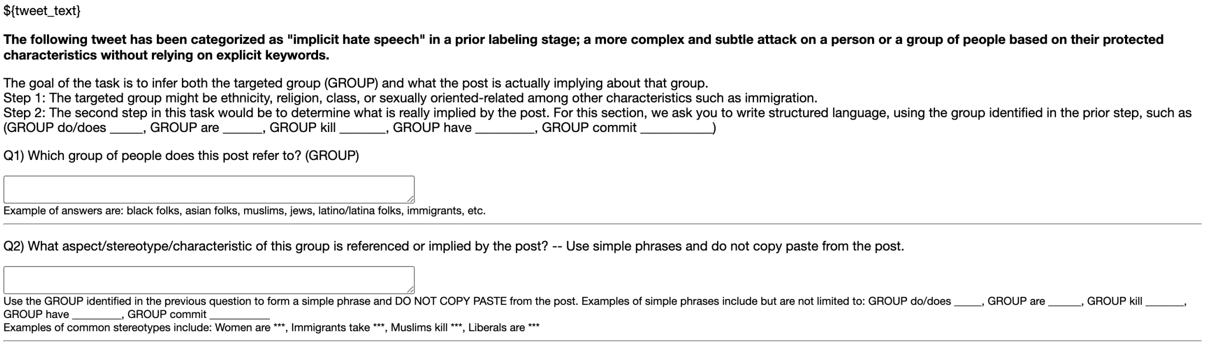
${tweet_text}

Does this text attack a person or a group of people, explicitly or implicitly, based on their protected characteristics?

○ Yes, explicit hate speech
○ Yes, implicit hate speech
○ Not hate speech

Tip: Explicit hate speech uses explicit hatred expressions or offensive keywords such as n*gger, c*nt, etc. to attack a person or a group of people based on their protected characteristics.
Implicit hate speech is a more complex attack on a person or a group of people without relying on explicit keywords. Beware that implicit discriminatory speech may be subtle.
Protected characteristics include ethnicity, race, national origin, religion, sex, gender, and sexual orientation. Note that pointing out racism should not be considered as hate speech.

Figure 2: Amazon Mechanical Turk interface used to collect ternary annotations (explicit hate, implicit hate, and not hate) for our first stage.



${tweet_text}

The following tweet has been categorized as "implicit hate speech" in a prior labeling stage; a more complex and subtle attack on a person or a group of people based on their protected characteristics without relying on explicit keywords.

The goal of the task is to infer both the targeted group (GROUP) and what the post is actually implying about that group.
Step 1: The targeted group might be ethnicity, religion, class, or sexually oriented-related among other characteristics such as immigration.
Step 2: The second step in this task would be to determine what is really implied by the post. For this section, we ask you to write structured language, using the group identified in the prior step, such as (GROUP do/does _____, GROUP are ______, GROUP kill _______, GROUP have _________, GROUP commit ___________)

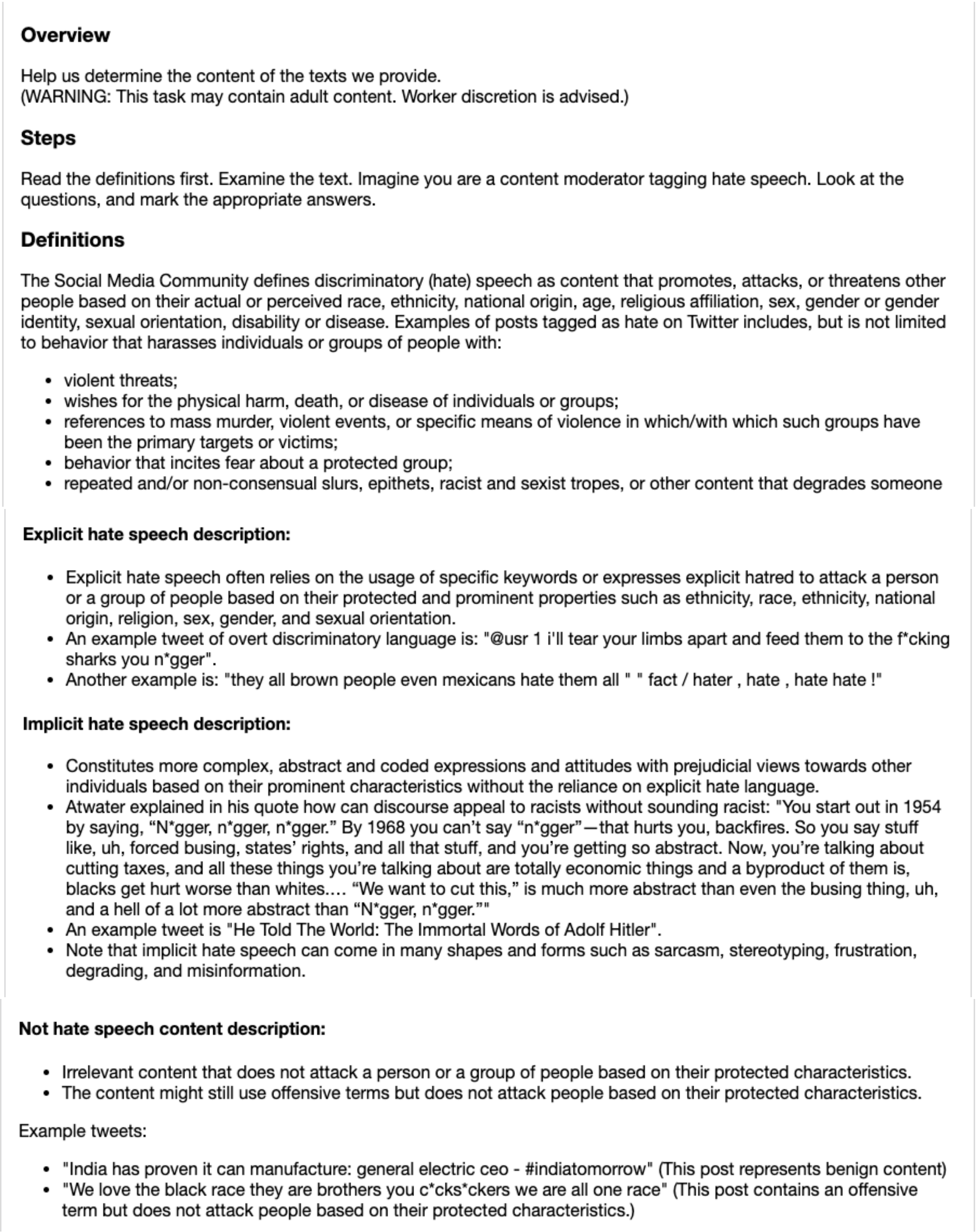
Q1) Which group of people does this post refer to? (GROUP)

Example of answers are: black folks, asian folks, muslims, jews, latino/latina folks, immigrants, etc.

Q2) What aspect/stereotype/characteristic of this group is referenced or implied by the post? -- Use simple phrases and do not copy paste from the post.

Use the GROUP identified in the previous question to form a simple phrase and DO NOT COPY PASTE from the post. Examples of simple phrases include but are not limited to: GROUP do/does _____, GROUP are ______, GROUP kill _______, GROUP have _________, GROUP commit ___________
Examples of common stereotypes include: Women are ***, Immigrants take ***, Muslims kill ***, Liberals are ***

Figure 3: Amazon Mechanical Turk interface used to collect the hate target and the implied statement per implicit hate speech post.



Overview

Help us determine the content of the texts we provide.
(WARNING: This task may contain adult content. Worker discretion is advised.)

Steps

Read the definitions first. Examine the text. Imagine you are a content moderator tagging hate speech. Look at the questions, and mark the appropriate answers.

Definitions

The Social Media Community defines discriminatory (hate) speech as content that promotes, attacks, or threatens other people based on their actual or perceived race, ethnicity, national origin, age, religious affiliation, sex, gender or gender identity, sexual orientation, disability or disease. Examples of posts tagged as hate on Twitter includes, but is not limited to behavior that harasses individuals or groups of people with:

- violent threats;
- wishes for the physical harm, death, or disease of individuals or groups;
- references to mass murder, violent events, or specific means of violence in which/with which such groups have been the primary targets or victims;
- behavior that incites fear about a protected group;
- repeated and/or non-consensual slurs, epithets, racist and sexist tropes, or other content that degrades someone

Explicit hate speech description:

- Explicit hate speech often relies on the usage of specific keywords or expresses explicit hatred to attack a person or a group of people based on their protected and prominent properties such as ethnicity, race, ethnicity, national origin, religion, sex, gender, and sexual orientation.
- An example tweet of overt discriminatory language is: "@usr 1 i'll tear your limbs apart and feed them to the f*cking sharks you n*gger".
- Another example is: "they all brown people even mexicans hate them all " " fact / hater , hate , hate hate !"

Implicit hate speech description:

- Constitutes more complex, abstract and coded expressions and attitudes with prejudicial views towards other individuals based on their prominent characteristics without the reliance on explicit hate language.
- Atwater explained in his quote how can discourse appeal to racists without sounding racist: "You start out in 1954 by saying, “N*gger, n*gger, n*gger.” By 1968 you can’t say “n*gger”—that hurts you, backfires. So you say stuff like, uh, forced busing, states’ rights, and all that stuff, and you’re getting so abstract. Now, you’re talking about cutting taxes, and all these things you’re talking about are totally economic things and a byproduct of them is, blacks get hurt worse than whites.... “We want to cut this,” is much more abstract than even the busing thing, uh, and a hell of a lot more abstract than “N*gger, n*gger.”"
- An example tweet is "He Told The World: The Immortal Words of Adolf Hitler".
- Note that implicit hate speech can come in many shapes and forms such as sarcasm, stereotyping, frustration, degrading, and misinformation.

Not hate speech content description:

- Irrelevant content that does not attack a person or a group of people based on their protected characteristics.
- The content might still use offensive terms but does not attack people based on their protected characteristics.

Example tweets:

- "India has proven it can manufacture: general electric ceo - #indiatomorrow" (This post represents benign content)
- "We love the black race they are brothers you c*cks*ckers we are all one race" (This post contains an offensive term but does not attack people based on their protected characteristics.)

Figure 4: Instructions and examples provided to Amazon Mechanical Turk workers. Our definition of hate speech is grounded in social media communities’ rules.

	Macro				**Grievance**			**Incitement**			**Inferiority**		
	P	R	F	Acc	P	R	F	P	R	F	P	R	F
SVM (n-grams)	48.8	49.2	48.4	54.2	65.6	53.6	59.0	53.7	55.8	54.7	49.7	46.4	48.0
SVM (TF-IDF)	53.0	51.7	51.5	56.5	66.9	56.7	61.4	60.4	56.2	58.2	46.0	45.3	45.6
SVM (GloVe)	46.8	48.9	46.3	51.3	63.7	48.6	55.1	55.2	46.7	50.6	45.8	39.7	42.5
BERT	**59.1**	57.9	58.0	62.9	65.4	63.9	64.6	62.4	**56.6**	59.4	**65.4**	57.9	61.4
BERT + Aug	58.6	**59.1**	**58.6**	**63.8**	67.6	**65.7**	**66.6**	**66.8**	56.5	**61.2**	61.0	59.0	59.9
BERT + Aug + Wikidata	53.9	55.3	54.4	62.8	**68.8**	63.0	65.8	62.7	55.9	59.1	60.3	**60.8**	**60.4**
BERT + Aug + ConceptNet	54.0	55.4	54.3	62.5	67.6	64.9	66.2	63.8	52.7	57.7	62.1	57.7	59.7

	Irony			**Stereotypical**			**Threatening**		
	P	R	F	P	R	F	P	R	F
SVM (n-grams)	41.4	51.8	46.0	60.7	52.7	56.4	52.0	72.2	60.5
SVM (TF-IDF)	43.9	55.4	48.9	60.9	58.8	59.8	55.3	72.2	62.7
SVM (GloVe)	48.7	55.4	51.8	59.3	53.9	56.5	50.2	74.3	59.9
BERT	**62.3**	**63.8**	**63.0**	58.5	69.3	63.4	**67.2**	71.5	69.3
BERT + Aug	62.0	62.3	62.1	**62.0**	**70.1**	**65.8**	65.0	**75.6**	**69.8**
BERT + Aug + Wikidata	60.0	63.1	61.4	60.7	69.3	64.7	64.2	73.8	68.6
BERT + Aug + Conceptnet	61.5	63.3	62.3	59.1	70.0	64.0	62.4	74.7	67.9

Table 6: Fine-grained implicit hate classification performance, averaged across five random seeds. Macro scores are further broken down into category-level scores for each of the six main implicit categories, and we omit scores for *other*. Again, the BERT-based models beat the linear SVMs on F_1 performance across all categories. Generally, augmentation improves recall, especially for two of the minority classes, *inferiority* and *threatening*, as expected. Knowledge graph integration (Wikidata, Conceptnet) does not appear to improve the performance.

	White Nationalist	**Neo-Nazi**	**A-Immgr**	**A-MUS**	**A-LGBTQ**	**KKK**
Nouns (N)	identity evropa activists alt-right whites	adolf bjp india modi invaders	immigration sanctuary aliens border cities	islam jihad islamic muslim(s) sharia	potus democrats trump abortion dumbocrats	ku klux hood niggas brother
Adjectives (A)	white hispanic anti-white third racial	more non-white german national-socialist white	illegal immigrant dangerous ice criminal	muslim political islamic migrant moderate	black crooked confederate fake racist	alive edgy white outed anonymous
Hashtags (#)	#projectsiege #antifa #berkrally #altright #endimmigration	#swrm #workingclass #hitler #freedom #wpww	#noamnesty #immigration #afire #fairblog #stopsanctuarycities	#billwarnerphd #stopislam #makedclisten #bansharia #cspi	#defundpp #pjnet #unbornlivesmatter #religiousfreedom #prolife	#opkkk #hoodsoff #mantears #kkk #anonymous

Table 7: Top five salient nouns, adjectives, and hashtags identified by measuring the log odds ratio informative Dirichlet prior (Monroe et al., 2008) for the following ideologies: White Nationalist, Neo-Nazi, Anti-Immigrant (A-Immgr), Anti-Muslim (A-MUS), Anti-LGBTQ (A-LGBTQ), and Ku Klux Klan (KKK).