Towards a Comprehensive Taxonomy and Large-Scale Annotated Corpus for Online Slur Usage

Jana Kurrek †
McGill University
School of Computer Science
jana.kurrek@mail.mcgill.ca

Haji Mohammad Saleem †
McGill University
School of Computer Science
haji.saleem@mail.mcgill.ca

Derek Ruths
McGill University
School of Computer Science
derek.ruths@mcgill.ca

Abstract

Abusive language classifiers have been shown to exhibit bias against women and racial minorities. Since these models are trained on data that is collected using keywords, they tend to exhibit a high sensitivity towards pejoratives. As a result, comments written by victims of abuse are frequently labelled as hateful, even if they discuss or reclaim slurs. Any attempt to address bias in keyword-based corpora requires a better understanding of pejorative language, as well as an equitable representation of targeted users in data collection. We make two main contributions to this end. First, we provide an annotation guide that outlines 4 main categories of online slur usage, which we further divide into a total of 12 sub-categories. Second, we present a publicly available corpus based on our taxonomy, with 39.8k human annotated comments extracted from Reddit. This corpus was annotated by a diverse cohort of coders, with Shannon equitability indices of 0.90, 0.92, and 0.87 across sexuality, ethnicity, and gender. Taken together, our taxonomy and corpus allow researchers to evaluate classifiers on a wider range of speech containing slurs.

1 Introduction

Detecting abusive language is important for two substantive reasons. First is the mitigation of harm to individuals. Exposure to hate speech can result in a wide range of psychological effects, including degradation of mental health, depression, reduced self-esteem, and greater stress expression (Saha et al., 2019; Tynes et al., 2008; Boeckmann and Liew, 2002). Second is the broader impact of unregulated speech on the participation gap in social media (Jenkins, 2009; Notley, 2009). Overexposure to hateful language results in user desensitization (Soral et al., 2018) and radicalization (Norman and Mikhael, 2017), both of which have been shown to worsen racial relations (Sène, 2019). Moreover, hateful echo-chambers promote a "spiral of silence" that discourages counter-speech in conversations online (Duncan et al., 2020).

Access to large-scale training data is the first step towards robust automated systems for abusive language detection. While industry researchers can access moderator logs and user reports, proprietary data is not the standard for academics. Instead, pejorative keywords are commonly used as filters in the data collection process. These include, but are not limited to, slurs and other curated lists of profane language (Waseem and Hovy, 2016; Waseem, 2016; Khodak et al., 2018; Rezvan et al., 2018), terms borrowed from Hatebase, a multilingual repository for hate speech (Silva et al., 2016; Davidson et al., 2017; Founta et al., 2018; ElSherief et al., 2018), offensive hashtags (Chatzakou et al., 2017; Golbeck et al., 2017), and manually selected threads or subreddits (Gao and Huang, 2017; Hammer et al., 2019; Qian et al., 2019). Although the drawbacks of keyword-based approaches are known to researchers, there are currently no clear alternatives to this technique (Waseem and Hovy, 2016; Davidson et al., 2017; ElSherief et al., 2018).

There has been a recent focus on how technical choices involving data curation can introduce systemic bias in the resultant corpus. For instance, Wiegand et al. (2019) discover that terms like *football*, *announcer*, and *sport* have the strongest correlation to abusive posts in Waseem and Hovy (2016). Furthermore, Davidson et al. (2019), Xia et al. (2020) and Sap et al. (2019) reveal how classifiers trained on data with systemic racial bias have a higher tendency to label text written in African-American English as abusive. Cited examples include: "Wussup, nigga!", and "I saw his ass yesterday". Left unaddressed, bias has a real impact on users. Automated recruiting tools

† These authors made equal contributions.

Proceedings of the Fourth Workshop on Online Abuse and Harms, pages 138–149
Online, November 20, 2020. ©2020 Association for Computational Linguistics
https://doi.org/10.18653/v1/P17

used by Amazon.com were shown to discriminate against women (Cook, 2018). Similarly, Microsoft released a public chatbot that learned to share racist content on Twitter (Vincent, 2016). A common solution is to debias language representations (Bolukbasi et al., 2016). However, these methods conceal but do not remove systemic bias in the overall data (Gonen and Goldberg, 2019).

A way of beginning to address the issue of racial and gender bias is therefore to understand the implications of forced sampling. Our paper focuses specifically on data that is collected using derogatory keywords and we make two main contributions to this end. First, we provide an annotation guide that outlines 4 main categories of online slur usage, which we further divide into a total of 12 subcategories. Second, we present a publicly available corpus based on our taxonomy, with 39.8k human annotated comments extracted from Reddit. We also propose an approach to data collection and annotation that prioritizes inclusivity both by design and application:

Inclusivity by Design: Data selection and annotation achieves weighted group representation. We sample from a variety of subreddits in order to capture non-derogatory slur usage. We then hire a diverse set of coders under strict ethical standards as a means of engaging the perspectives of various target communities. We encourage opinion diversity by pairing annotators into teams based on maximum demographic differences.

Inclusivity by Application: Our coding guidelines are extensible to language that targets multiple protected groups. We collect data using the slurs: *faggot*, a pejorative term used primarily to refer to gay men, *nigger*, an ethnic slur typically directed at black people, especially African Americans, and *tranny*, a derogatory slur for a transgender person. This is only time we mention the actual slurs. From hereon, We refer to each term as the f-slur, n-slur, and t-slur, respectively. We specifically choose these slurs because they enable us to study discrimination across sexuality, ethnicity, and gender.

Our work does not directly eliminate bias in existing datasets. Rather, it aids in truly understanding the different ways in which slurs can be used online so that models can be trained and assessed more effectively.

2 Related Work

2.1 Existing Hate Speech Corpora

The earliest and most notable corpus for hate speech research is Waseem and Hovy (2016). It contains 16k comments from Twitter, annotated according to the offense criteria of McIntosh (1988). Waseem (2016) is an extension of this corpus by 6,909 comments and it considers amateur as well as expert annotations. The authors make use of offensive hashtags for data collection, but it was not until Nobata et al. (2016) that slurring language was formally introduced as a sub-problem of hate speech. This paper uses a variety of linguistic features, such as modal words, insulting and hate blacklist words, and politeness words, in order to separate the three notions of hate, derogation, and profanity based on their relative degrees of harm to the target. These guidelines inspired the Fox News user comments corpus of Gao and Huang (2017). Both works emphasize the capacity for hateful language to exist in implicit and explicit forms and collect the explicit form using derogatory keywords. Silva et al. (2016) is a target-based analysis of the explicit form. They leverage the syntactic structure "`I <intensity><user intent><hate target>`", where each hate target is one of 1,078 terms selected from Hatebase, in order to identify ten top targets of hate within Twitter and Whisper content. Next, Davidson et al. (2017) investigate intentional group-based humiliation and derogation. They reinforce the role of slurs as archetypal representations of hate by acknowledging that "tweets with the highest predicted probabilities of being hate speech tend to contain multiple racial or homophobic slurs." More recently, de Gibert Bonet et al. (2018) sample from a white supremacist sub-forum and, in doing so, encourage community-based filtering. The emerging theme from these research efforts is the consensus that we require an alternative to random sampling for reliably capturing hateful content. What that alternative is remains unclear but keywords are currently the dominant choice.

Other researchers have expanded on this definition and shown that it is applicable to more nuanced categories of online misbehaviour, such as abuse, threats, personal attacks, and cyberbullying. For instance, Khodak et al. (2018) is a self-annotated corpus for sarcasm on Reddit. Sprugnoli et al. (2018) focuses on cyberbullying within WhatsApp conversations. Rezvan et al. (2018) points out sexual,

Authors	Size	Platform	Annotation	Agreement
KEYWORD BASED DATA COLLECTION				
Qian et al. (2019)	34k	Gab	Hate Speech (Binary)	Unknown
Qian et al. (2019)	22k	Reddit	Hate Speech (Binary)	Unknown
Waseem and Hovy (2016)	16k	Twitter	Racism, Sexism	κ = 0.84
Waseem (2016)	7k	Twitter	Racism, Sexism	κ = 0.34 (Majority Vote) κ = 0.70 (Full Agreement)
Golbeck et al. (2017)	35k	Twitter	Hate Speech, Threats, Harassment, Offense	κ = 0.84
Chatzakou et al. (2017)	9k	Twitter	Aggressors, Bullies, Spammers	Inter-rater agreement = 0.54
Davidson et al. (2019)	25k	Twitter	Hate Speech, Offense	Inter-rater agreement = 0.92
Rezvan et al. (2018)	25k	Twitter	Harassment	κ = 0.70; 0.84; 1.0; 0.80; 0.69 for respective categories
Founta et al. (2018)	80k	Twitter	Hate Speech, Spam, Abuse	Unknown
ElSherief et al. (2018)	2k	Twitter	Hate Speech	α = 0.622
Jha and Mamidi (2017)	1k	Twitter	Sexism	$F\kappa$ = 0.74
Silva et al. (2016)	539.5m	Twitter Whisper	Hate Speech	Not applicable
Fersini et al. (2018)	3k	Twitter	Sexism	Unknown
Basile et al. (2019)	19.6k	Twitter	Hate Speech, Target, Aggressiveness	F8 confidence = 0.83 0.70, 0.73
Zampieri et al. (2019)	14.1k	Twitter	Offense, Target	$F\kappa$ = 0.83* *on 21 tweets
MANUAL SELECTION				
Gao and Huang (2017)	1.5k	Fox News	Hate Speech	κ = 0.98
Hammer et al. (2019)	30k	Youtube	Threats	Unknown
PROPRIETARY DATA				
Sprugnoli et al. (2018)	15k	WhatsApp	Cyberbullying	SDC = 0.80 - 0.88
Nobata et al. (2016)	1.2m	Yahoo	Hate Speech	$F\kappa$ = 0.40; 0.21 for AMT $F\kappa$ = 0.84; 0.46 for Trained (Binary; Fine-grained)
RANDOM DATA SELECTION				
de Gibert Bonet et al. (2018)	10k	Stormfront	Hate Speech (Binary)	κ = 0.61; $F\kappa$ = 0.61 (Batch1) κ = 0.63; $F\kappa$ = 0.63 (Batch2)
Napoles et al. (2017)	10k	Yahoo	Positive Conversations	α = 0.79 (Group) α = 0.71 (Trained)
OTHER METHODS				
Wulczyn et al. (2017)	100k	Wikipedia	Harassment, Attacks	α = 0.45
Kennedy et al. (2017)	20k	Twitter, Reddit, The Guardian,	Harassment (Binary)	Inter-rater agreement = 0.88

Table 1: An overview of the main corpora on abusive language and similar behaviours. $F\kappa$ is Fleiss' Kappa, κ is Cohen's Kappa, SDC is the Sørensen–Dice coefficient, and inter-rater agreement refers to raw disagreement.

appearance-related, intellectual, and political harassment on Twitter. Hammer et al. (2019) is a corpus for detection of violent threats on YouTube. Holgate et al. (2018), Cachola et al. (2018), and Pamungkas et al. (2020) examine vulgarity and swearing. A number of corpora on mixed behaviours have also been produced. Golbeck et al. (2017) is a study on harassment and offense on Twitter. Chatzakou et al. (2017) labels Twitter users, not comments, as aggressors, bullies, or spammers. Founta et al. (2018) considers spam in conjunction with abuse, bullying, and aggression on Twitter. Napoles et al. (2017) works on the converse of the problem. This paper uses Yahoo! News data to advance a corpus on constructive conversations.

We have collected a list of the major English-language corpora and summarized their sizes, platforms of focus, annotation schemes, and agreement scores in Table 1. With that said, the study of online misbehavior has been extend beyond the traditional focus on English. It now includes resources in Italian, Indonesian, Hindi-English, Tunisian, etc. (Sanguinetti et al., 2018; Ibrohim and Budi, 2018; Kumar et al., 2018; Bohra et al., 2018; Haddad et al., 2019; Mulki et al., 2019; Chung et al., 2019).

2.2 Slurs

To model the contents of slur-based data, it is crucial that we first examine the properties of slurs themselves. Slurs are pejoratives that derogate based on in-group membership, that is, they categorize targets based on institutionally defined

f-slur	n-slur	t-slur
SUPPORTIVE COMMUNITIES		
askgaybros	BlackPeopleTwitter	transgendercirclejerk
gaybros	Blackfellas	traaaaaaannnnnnnnns
lgbt	blackladies	asktransgender
ainbow	beholdthemasterrace	ainbow
LGBTeens	AgainstHateSubreddits	transgender
ANTAGONISTIC COMMUNITIES		
4chan	CoonTown	TumblrInAction
ImGoingToHellForThis	uncensorednews	MGTOW
The_Donald	WhiteRights	Braincels
CringeAnarchy	GreatApes	metacanada
TheRedPill	european	GenderCritical
GENERAL DISCUSSION COMMUNITIES		
funny	todayilearned	rupaulsdragrace
pics	videos	cars
politics	changemyview	Drama
AskReddit	worldnews	AdviceAnimals
atheism	movies	unpopularopinion

Table 2: This table presents the major supportive, antagonistic, and general discussion subreddits that were used in data collection. Their range of views towards the targets of each slur facilitates equitable representation.

archetypes (Croom, 2015). Studies on slurs are built on the recognition by Kaplan (1999) that meaning in natural language comes from convention and from context: a sentence is *expressively correct* if it is true by interpretation; a sentence is *descriptively correct* if it is literally true.

Hom (2008) advocates in favor of the expressive view of slurs. He identifies nine adequacy conditions that characterize and explain racial epithets: A slur exhibits (1) derogatory force. The force of any slur is (2) variable across epithets and (3) fundamentally offensive, independently of the intents and beliefs of the speaker. While slurs are capable of being (7) reclaimed or (8) used towards a non-derogatory, non-appropriative end, they are generally (4) taboo unless (6) their force changes over time. This is because slurs are (5) meaningful insofar as they contribute to the truth-conditions of the sentence in which they arise. Hom's account of slurs is (9) generalizable across pejoratives.

Hom implies that there are three main categories of slur usage, which are derogatory, non-derogatory non-appropriative, and appropriative. His adequacy conditions are central to our research. The three categories are the basis of our annotation scheme and they enable us to make assessments of abuse with ambiguous user intent.

3 Inclusive Design Process

Random sampling of slur-based data allows for proportional representation because the share of each usage in the corpus is reflective of its probability of occurrence online. However, this approach is not equitable. Less common usages, such as reclamation, discussion, and counter-speech, are not captured. Consequently, language models can overfit on pejoratives and further codify institutional biases (Caliskan et al., 2017; Garg et al., 2018). A top-down approach to debiasing is simply insufficient. We advocate in favor of affirmative action during data collection and make an effort to represent a wider range of slur usages through community targeting. We also tailor our study to include individuals that belong to targeted communities, both as authors and annotators.

3.1 Data Collection

We use the Pushshift Reddit corpus (Baumgartner et al., 2020) and filter for the three slurs (f-slur, n-slur, t-slur) and their plurals. The data ranged from October 2007 to September 2019 at the time of filtering. We extracted a total of 2.6 million comments. We applied the following filtering process:

Author Level: We remove comments written by users with no history in order to leave open the possibility of a future analysis with user meta-data. We remove comments written by users that were identified as bots. We limit the number of comments written by the same author.

Comment Level: Reddit comments vary in length, with an upper limit of 40,000 characters. For ease of annotation, we remove comments from the top and bottom quartiles by length. We limit our corpus to English-language comments and use the

Slur Usages	Example Text
DEROGATORY	
Attribution	he's an ugly [f-slur] with greasy hair.
Community Focus	lol don't be a [f-slur]
Stand Alone	[t-slur]
Sexualization	I love the taste of a nice hot [t-slur] load
Self-Deprecation	as mizkif i can agree i look like a [f-slur]
APPROPRIATIVE	
Reclamation	get in [t-slur] Formation everyone, it's time to march against the tyranny of heteronormatives trying to appropriate OUR WORDS
NON-DEROGATORY, NON-APPROPRIATIVE	
Counter Speech	[t-slur] is a slur please don't use it.
Direct Quotations	actual quote: de [n-slur] woman is de mule uh de world so fur as ah can see.
Discussion	You could call someone a [f-slur] in the 70s and 80s with absolutely no recourse.
Recollection	I never got so much shit until I graduated high school. :— I get called a [f-slur] by some random clitdick almost every day I have class.
Sarcasm	Yeah because apparently [f-slur] all of a sudden isn't a slur used against homosexuals.
HOMONYMS	
	transmissions are beautiful pieces of engineering, why not have a [t-slur] tattoo? [f-slur] Hill, 969th tallest peak in Massachusetts... why even count at that point? Damn talk about being able to skate anything. Rips [t-slur] then throws in kickflip back lips on rails.

Table 3: Our taxonomy of slur usage, with 4 main categories broken down into 12 subcategories. Examples are provided for each subcategory and further detail can be found in the Appendix.

Compact Language Detector v3[1] to detect them.

Community Level: Communities that antagonise or support a group talk about similar topics but with opposing valence (Saleem et al., 2016). To capture such polarity, we compile a list of subreddits based on their disdain for, neutrality towards, or support of the f-slur, n-slur, and t-slur (see Table 2). We do this by building on an existing list of toxic Reddit communities (Caffier, 2017). We consider the name, rules, extent of moderation, description text, and polarity of comments containing slurs (overall score) of each subreddit in our assessment of whether or not to include them. We then extract the top comments in terms of polarity.

Our post-filter corpus has 40,000 comments, sourced from 2704 individual subreddits and 37,133 unique authors. The median and maximum number of comments per author is 1 and 5.

3.2 Taxonomy Design

Our coding guide is based on the three major categories of slur usage identified in Hom (2008). By open coding data collected using slurs, we identify a fourth major category as well as twelve subcategories. The complete taxonomy, along with examples for each subcategory, is provided in Table 3. In general, comments containing more than one slur were labelled according to the most derogatory usage. The four main categories are explained below:

Derogatory Usage (DER): Any usage that is understood to convey contempt towards a targeted individual or group.

Appropriative Usage (APR): Meaningful usage by the targeted group for an alternate, non-derogatory purpose. Text belonging to this label loses its derogatory force.

Non-Derogatory, Non-Appropriative Usage (NDG): Meaningful usage by targeted or non-targeted groups for an alternate non-derogatory, non-appropriative purpose. Text belonging to this label retains its derogatory force.

Homonyms (HOM): A slur with one or more non-derogatory alternative meanings.

3.3 Annotator Selection

Following approval by the university Research Ethics Board (REB), we shared messages on social media and university mailing lists as well as physical posters across faculties in order to look for participants. The application consisted of eight short answer questions, in which candidates were asked to disclose their name, email, field and year of study, age, sexuality, ethnicity, and gender. We specifically collected the demographic information in free-form text. The free-form allows participants to choose best demographic identifiers for themselves. The demographic information is confidential and used solely for selecting annotators and creating their teams.

[1] https://github.com/google/cld3

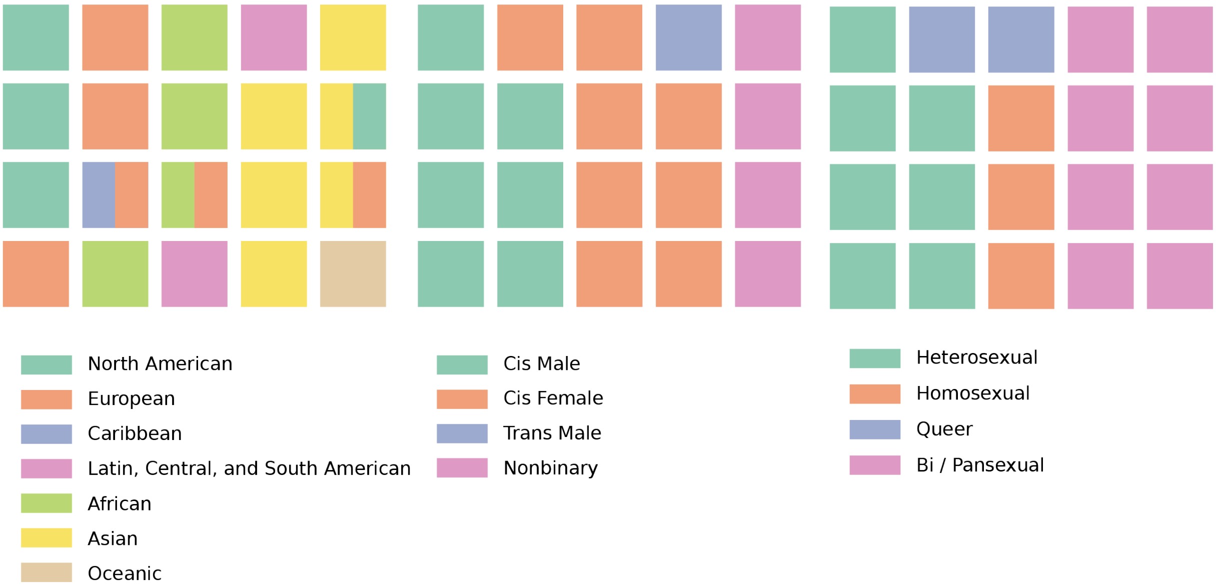

Figure 1: The diverse demographic details of our annotator cohort, aggregated on ethnicity, gender and sexuality.

All demographics were collapsed into categories (see Figure 1) primarily based on the classification structure approved as a departmental standard by Statistics Canada (2017). Of the four hundred and twelve applications received, 20 participants, ranging between 19 and 65 years of age (M = 26.7, SD = 10.8), were chosen using iterative proportional fitting. Overall, our annotator cohort has a Shannon equitability index of 0.90, 0.92, and 0.87 across sexuality, ethnicity, and gender. We did not have the REB clearance to perform any further analysis on the relationship between annotator demographics and annotations. We leave this as an area for future work.

3.4 Training and Annotation

A 4-session on-campus training program was developed for annotators to attend over 2 days. On Day 1, we presented the annotation scheme obtained through open coding. Annotators were then guided through two group annotation exercises of 20 and 40 comments respectively. On Day 2, annotators were randomly divided into 4 teams. Each team completed 2 rounds of 200 training annotations. After each round, they discussed their annotations and the reasons behind their labels. The discussion was aimed at fostering a common understanding of the annotation process.

The final annotations were divided into 4 tasks of 10,000 comments each. The 20 annotators were grouped into 10 teams of 2. The team creation process maximized the demographic distance between members across sexuality, ethnicity, and gender. It was treated as an assignment problem and solved using the Kuhn-Munkres algorithm. Each team annotated 1000 comments per task and annotators were grouped into new pairs for each subsequent task. Comments with no disagreement were added to the final corpus. Comments with disagreement were resolved by the authors. The final annotations were performed remotely on the open source text annotation tool Doccano (Nakayama et al., 2018).

4 Labeled Corpus

40,000 Reddit comments were annotated, of which 189 were removed as noise. The remaining 39,811 were closely split across slurs: 13,290, 13,267 and 13,267 for f-slur, n-slur and t-slur respectively. In total, 20,531 comments were labelled derogatory, 16,729 non-derogatory, 1,998 homonym, and 553 appropriative. We anticipated a large portion of derogatory comments in our corpus because our data is slur-based. However, only 52% of comments were labelled as such. We attribute this to our community-targeted data collection process and efforts to sample from supportive subreddits.

4.1 Label Distribution Across Slurs

In Figure 2, we present the label distribution across slurs. We observe that roughly 59% of comments collected using the f-slur and t-slur were labelled as derogatory. In comparison, about 37.9% of comments containing the n-slur were similarly labelled. The majority of found homonyms include the t-

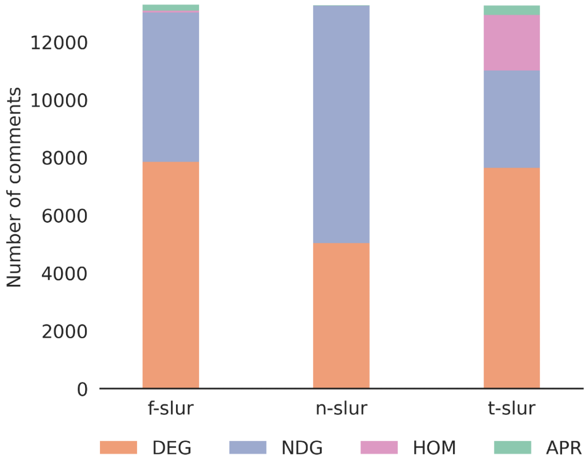

Figure 2: The label distribution across slurs.

slur, which accounts for 95.9% of the label. This is largely because the term is used in automotive communities to mean vehicle transmission (see Figure 3) and in skateboarding communities to describe skating transition. The remaining homonyms include the f-slur, with the meaning "bundle" or in reference to a form of British meatball. The n-slur has the smallest share of homonyms (0.02%) and appropriative (0.16%) comments.

4.2 Label Distribution Across Subreddits

In Figure 3, we present the label distribution across the 50 most common subreddits in our corpus. The graph is sorted by the proportion of derogatory comments in each subreddit. Consequently, it can be seen as a scale of derogatory behavior. On the far right are communities that we had previously identified as antagonistic. Many of their comments were labelled as derogatory and examples include `MGTOW`, `CoonTown`, `4chan` and, `The_Donald`. In the middle we find general discussion subreddits such as `videos`, `todayilearned`, and `politics`. They generally have an even split of derogatory and non-derogatory labels. On the far left we observe mostly supportive subreddits, with small portions of derogatory comments. Automotive subreddits like `cars` have a large number of homonyms. Meanwhile, subreddits such as `traaaaaaannnnnnnnnns`, `askgaybros`, and `rupaulsdragrace` contain significant portion of appropriative speech. These findings align with our initial hypothesis about supportive, antagonistic, and general discussion communities.

	Agreement (%)	Cohen's κ
overall	78.6	0.60
f-slur	79.7	0.58
n-slur	75.4	0.51
t-slur	80.5	0.65

Table 4: Raw and inter-rater agreement. We achieve moderate to substantial agreement with Cohen's κ.

4.3 Agreement Analysis

Both annotators agreed on the same label for 31,034 of the comments in our corpus. The remaining 8,777 comments were resolved by the authors. Overall we achieve a raw agreement score of 78.6%, corresponding to a Cohen's κ of 0.60. Our scores indicate substantial agreement and are in line with what has been observed in the literature (see Table 1). We obtain similar agreement across the three slurs, which are presented in Table 4.

APR had the highest amount of disagreement, with 67.99% comments requiring resolution, followed by NDG (35.36%), and HOM (31.58%). DEG was the lowest at 9.034%. During the resolution process, we identified three probable causes for disagreement:

Label Overlap Discussions of derogation or reclamation created ambiguity and were falsely labelled as DER or APR, rather than NDG. A similar issue arose in comments acknowledging slurs as homonyms. For instance: *"When i was telling my skate friends about me being trans i asked them if they knew why it was so ironic that i love skating [t-slur] so much."*.

Satire Our annotators found many derogatory comments in `transgendercirclejerk` (see Figure 3), which is a subreddit that self-identifies as a "parody for trans people, mocking all transgender-related topics". However, the sarcastic or satirical nature of these comments was not always evident: *"We don't need gun control we need [T-SLUR] CONTROL! [t-slurs] are not in the Constitution or Bible, like guns are! If we don't outlaw t-slurs, only [t-slurs] will have outlaws!"*. We leave this area for future work.

Lack of Context In an independent assessment of label reliability, we re-annotated 100 DEG comments from `transgendercirclejerk` with complete access to user and thread history. 44 of our labels did not match those submitted by annotators. For instance, the following comment came from a transgender poster: *"LA LA LA CAN'T*

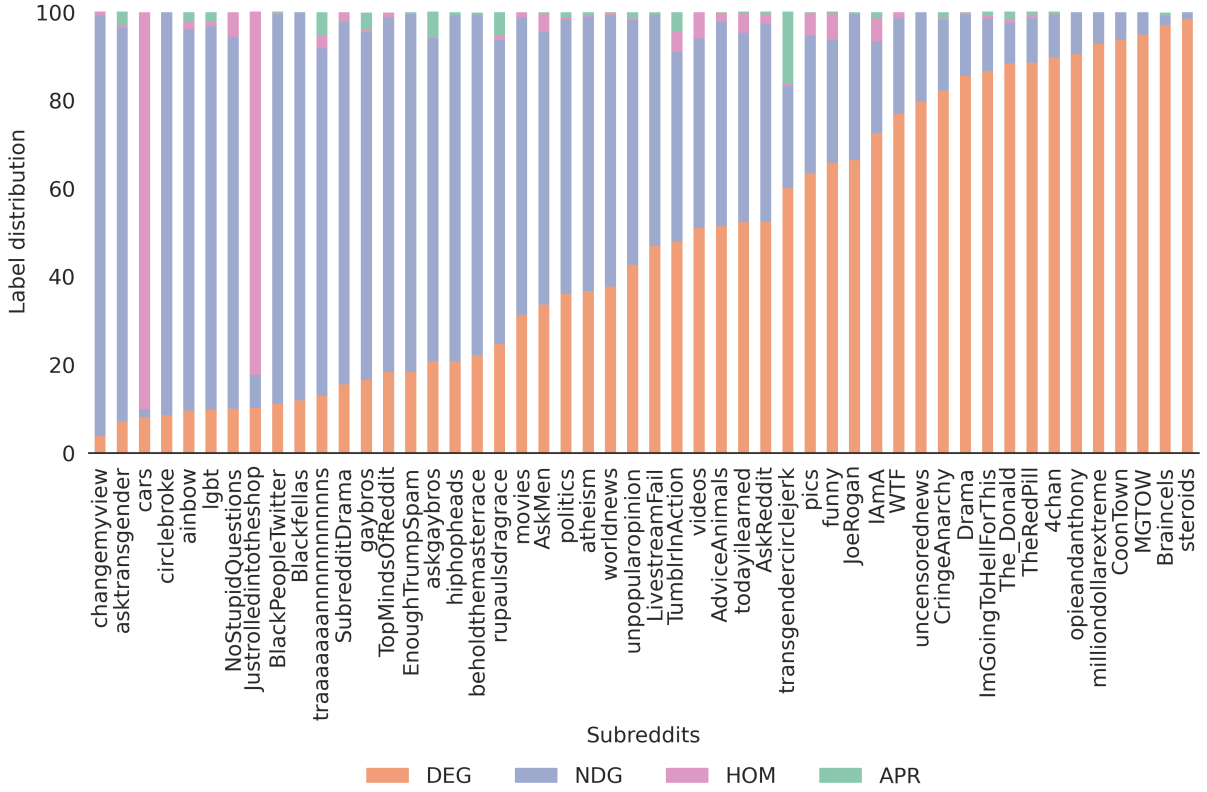

Figure 3: The normalized label distribution across the 50 most common subreddits in our corpus, sorted by their portion of derogatory comments.

HEAR YOU I'M STUCK IN [T-SLUR] REALITY" but was mislabelled. This testifies the difficulty of annotating appropriative language without context. Other instances that requires context are reference to lyrics and dialogues from pop culture.For example "Dead [n-slur] Storage" from the movie Pulp Fiction.

4.4 Benchmarking the Perspective API

We use a state-of-the-art model for derogatory content detection to assess whether current classifiers are subject to overfitting on pejoratives. We choose the Perspective API by Conversation AI, which "identifies whether a comment could be perceived as toxic to a discussion". We obtain the toxicity scores for 100 random comments for each of the DEG, NDG, HOM, and APR labels. The results are summarized in Figure 4. As expected, the overall score distribution is high for DEG. However, it is equally high for NDG and APR comments. This perfectly illustrates the issue of potentially biased models failing to identify non-derogatory content.

Further analysis of toxicity scores across comments underlines the challenges faced by existing models. First, instances of slur reclamation received high toxicity scores. For example: *"Psh my [t-slur] agony sits atop that steed with militant fervour. The world shall hear me roar, I AM A [T-SLUR] FREAK!!!! /uj Not even kidding, I'm 100% out as a [t-slur] freak. World can suck my shenis"* and *"When I've got a guy I'm crushing on I will sometimes say 'He makes me feel like a*

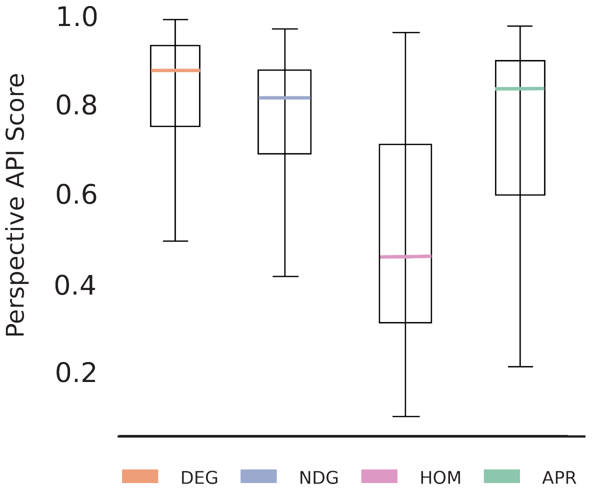

Figure 4: Benchmarking the Perspective API. Scores indicate a comment's degree of toxicity.

silly [f-slur] all over again'" have toxicity scores above 0.93. Reclamation is an attempt at empowerment and community cohesion. The mislabelling of such examples further censors communities already targeted by hate. Second, recollections of past harassment received high toxicity scores. For example: *"A homeless dude called me a spic [f-slur] once while I was with my ex"* is rated as high as 0.889. This belittles victims' experiences with abuse, rather than protecting them from it. Finally, counter speech received high toxicity scores. For example: *"Ummmm, yeah no, [t-slur] is a slur and youre ignorant as hell"* is rated 0.953. This undermines community-level efforts at removing derogatory language. Overall, these three outcomes are counterproductive to the detection process since empowering and vulnerable conversations of targeted communities may be flagged down.

5 Conclusion

We present a comprehensive taxonomy and large-scale annotated corpus for online slur usage. Our findings are an attempt at integrating a qualitative understanding of slurs into their usage in natural language. We believe that they provide a significant contribution to the hate speech research community, not only as resources for training machine and deep learning models, but also as a means of achieving a nuanced understanding of the phenomenon of slurs. We encourage researchers to replicate and expand our efforts by studying language that targets other marginalized communities. With that said, our corpus is a challenging benchmark that will help expose over-fitting on pejoratives and our taxonomy introduces a systematic approach for dealing with derogatory keywords and epithets. Our corpus can be accessed by emailing the authors.

References

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.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. *Proceedings of the International AAAI Conference on Web and Social Media*, 14(1):830–839.

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

Aditya Bohra, Deepanshu Vijay, Vinay Singh, Syed Sarfaraz Akhtar, and Manish Shrivastava. 2018. A dataset of hindi-english code-mixed social media text for hate speech detection. In *Proceedings of the second workshop on computational modeling of people's opinions, personality, and emotions in social media*, pages 36–41.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*, pages 4349–4357.

Isabel Cachola, Eric Holgate, Daniel Preoţiuc-Pietro, and Junyi Jessy Li. 2018. Expressively vulgar: The socio-dynamics of vulgarity and its effects on sentiment analysis in social media. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2927–2938, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Justin Caffier. 2017. Here are reddit's whiniest, most low-key toxic subreddits. *Vice.com*.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Elisabeth Camp. 2012. Sarcasm, pretense, and the semantics/pragmatics distinction. *Noûs*, 46(4):587–634.

Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. 2017. Mean birds: Detecting aggression and bullying on twitter. In *Proceedings of the 2017 ACM on web science conference*, pages 13–22.

Yi-Ling Chung, Elizaveta Kuzmenko, Serra Tekiroglu, and Marco Guerini. 2019. Conan - counter narratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2819–2829.

James Cook. 2018. Amazon scraps 'sexist ai' recruiting tool that showed bias against women. *The Telegraph*.

Adam M Croom. 2015. The semantics of slurs: A refutation of coreferentialism. *Ampersand*, 2:30–38.

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. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Eleventh international aaai conference on web and social media*.

Megan Duncan, Ayellet Pelled, David Wise, Shreenita Ghosh, Yuanliang Shan, Mengdian Zheng, and Doug McLeod. 2020. Staying silent and speaking out in online comment sections: The influence of spiral of silence and corrective action in reaction to news. *Computers in Human Behavior*, 102:192–205.

Mai ElSherief, Vivek Kulkarni, Dana Nguyen, William Yang Wang, and Elizabeth Belding. 2018. Hate lingo: A target-based linguistic analysis of hate speech in social media. In *Twelfth International AAAI Conference on Web and Social Media*.

Elisabetta Fersini, Paolo Rosso, and Maria Anzovino. 2018. Overview of the task on automatic misogyny identification at ibereval 2018. In *IberEval@ SEPLN*, pages 214–228.

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. In *Twelfth 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.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Ona de Gibert Bonet, Naiara Perez Miguel, 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. Association for Computational Linguistics.

Jennifer Golbeck, Zahra Ashktorab, Rashad O Banjo, Alexandra Berlinger, Siddharth Bhagwan, Cody Buntain, Paul Cheakalos, Alicia A Geller, Rajesh Kumar Gnanasekaran, Raja Rajan Gunasekaran, et al. 2017. A large labeled corpus for online harassment research. In *Proceedings of the 2017 ACM on web science conference*, pages 229–233.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. 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 609–614.

Philip B Gove. 1964. Noun often attributive" and" adjective. *American Speech*, 39(3):163–175.

Hatem Haddad, Hala Mulki, and Asma Oueslati. 2019. T-hsab: A tunisian hate speech and abusive dataset. In *International Conference on Arabic Language Processing*, pages 251–263. Springer.

Hugo L Hammer, Michael A Riegler, Lilja Øvrelid, and Erik Velldal. 2019. Threat: A large annotated corpus for detection of violent threats. In *2019 International Conference on Content-Based Multimedia Indexing (CBMI)*, pages 1–5. IEEE.

Eric Holgate, Isabel Cachola, Daniel Preoţiuc-Pietro, and Junyi Jessy Li. 2018. Why swear? analyzing and inferring the intentions of vulgar expressions. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4405–4414, Brussels, Belgium. Association for Computational Linguistics.

Christopher Hom. 2008. The semantics of racial epithets. *The Journal of Philosophy*, 105(8):416–440.

Muhammad Okky Ibrohim and Indra Budi. 2018. A dataset and preliminaries study for abusive language detection in indonesian social media. *Procedia Computer Science*, 135:222–229.

Henry Jenkins. 2009. *Confronting the challenges of participatory culture: Media education for the 21st century*. Mit Press.

Akshita Jha and Radhika Mamidi. 2017. When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data. In *Proceedings of the second workshop on NLP and computational social science*, pages 7–16.

David Kaplan. 1999. The meaning of ouch and oops. explorations in the theory of meaning as use. University of California.

George Kennedy, Andrew McCollough, Edward Dixon, Alexei Bastidas, John Ryan, Chris Loo, and Saurav Sahay. 2017. Technology solutions to combat online harassment. In *Proceedings of the first workshop on abusive language online*, pages 73–77.

Mikhail Khodak, Nikunj Saunshi, and Kiran Vodrahalli. 2018. A large self-annotated corpus for sarcasm. In *Proceedings of the Linguistic Resource and Evaluation Conference (LREC)*.

Ritesh Kumar, Aishwarya Reganti, Akshit Bhatia, and Tushar Maheshwari. 2018. Aggression-annotated corpus of hindi-english code-mixed data. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA).

Peggy McIntosh. 1988. White privilege: Unpacking the invisible knapsack.

Hala Mulki, Hatem Haddad, Chedi Bechikh Ali, and Halima Alshabani. 2019. L-hsab: A levantine twitter dataset for hate speech and abusive language. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 111–118.

Hiroki Nakayama, Takahiro Kubo, Junya Kamura, Yasufumi Taniguchi, and Xu Liang. 2018. doccano: Text annotation tool for human. Software available from https://github.com/doccano/doccano.

Courtney Napoles, Aasish Pappu, and Joel Tetreault. 2017. Automatically identifying good conversations online (yes, they do exist!). In *Eleventh International AAAI Conference on Web and Social Media*.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, pages 145–153.

Julie Norman and Drew Mikhael. 2017. Youth radicalization is on the rise. here's what we know about why. *The Washington Post*.

Tanya Notley. 2009. Young people, online networks, and social inclusion. *Journal of Computer-Mediated Communication*, 14(4):1208–1227.

Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2020. Do you really want to hurt me? predicting abusive swearing in social media. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6237–6246, Marseille, France. European Language Resources Association.

Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Wang. 2019. A benchmark dataset for learning to intervene in online hate speech. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4757–4766. Association for Computational Linguistics.

Mohammadreza Rezvan, Saeedeh Shekarpour, Lakshika Balasuriya, Krishnaprasad Thirunarayan, Valerie L Shalin, and Amit Sheth. 2018. A quality type-aware annotated corpus and lexicon for harassment research. In *Proceedings of the 10th ACM Conference on Web Science*, pages 33–36.

Koustuv Saha, Eshwar Chandrasekharan, and Munmun De Choudhury. 2019. Prevalence and psychological effects of hateful speech in online college communities. In *Proceedings of the 10th ACM Conference on Web Science*, pages 255–264.

Haji Mohammad Saleem, Kelly P Dillon, Susan Benesch, and Derek Ruths. 2016. A web of hate: Tackling hateful speech in online social spaces.

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.

Yaye Nabo Sène. 2019. Hate speech exacerbating societal, racial tensions with 'deadly consequences around the world', say un experts. *UN News*.

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

Wiktor Soral, Michał Bilewicz, and Mikołaj Winiewski. 2018. Exposure to hate speech increases prejudice through desensitization. *Aggressive behavior*, 44(2):136–146.

Rachele Sprugnoli, Stefano Menini, Sara Tonelli, Filippo Oncini, and Enrico Piras. 2018. Creating a whatsapp dataset to study pre-teen cyberbullying. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 51–59.

Robert Truswell. 2004. *Attributive adjectives and the nominals they modify*. Ph.D. thesis, Citeseer.

Brendesha M Tynes, Michael T Giang, David R Williams, and Geneene N Thompson. 2008. Online racial discrimination and psychological adjustment among adolescents. *Journal of adolescent health*, 43(6):565–569.

James Vincent. 2016. Twitter taught microsoft's ai chatbot to be a racist asshole in less than a day. *The Verge*, 24.

Zeerak Waseem. 2016. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the first workshop on NLP and computational social science*, pages 138–142.

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.

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.

Deirdre Wilson. 2006. The pragmatics of verbal irony: Echo or pretence? *Lingua*, 116(10):1722–1743.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1391–1399.

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.