An Annotation Guide for Online Slur Usage

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.

Derogatory Slurs

Derogatory slur usage is any usage that is understood to convey contempt towards a targeted individual or group [3]. It is divisible into the following five subcategories:

- 1. Attribution
- 2. Community Focused Slurs
- 3. Stand-Alone Slurs
- 4. Sexualization
- 5. Self-Deprecation

Attribution In general, a slur may be used as an attributive noun or as an attributive adjective. An attributive adjective typically follows an Adjective-Noun structure. Its structure can be recursed any number of times within a noun phrase, but attributive order and scope may affect the meaning of the sentence [4]. An attributive noun follows a free-form or bound-form compound structure [2] and slurring adjectives may be used as the base for noun inflections [5]. Examples of noun inflections are: $ni^{***ring}$, $tr^{***iversary}$, and f^{**qy} .



Figure 1. Noun-Noun Structure



Figure 2. Adj-Noun Structure



Figure 3. Adj-Noun-Noun Structure

Implicit attributive nouns target an entity that is unnamed but present in the sentence. This entity has attributes that are conventionally implicated by the slur. Such is the case with the n-slur in n-slur, I'm tired of this. Two statements are made by the speaker: the first, (you are the n-slur), is implicit, and the second, (I'm tired of this), is explicit. The sentence is therefore semantically equivalent to (you are the n-slur) & (I'm tired of this). Similarly, in not get downvoted by every f-slur on reddit for everything I say, f-slur acts as an implicit attributive noun because it suggests that (everyone who downvotes me is the f-slur). Other examples of implicit attributive nouns are nicknames and informal vocatives [2].

Community Focused Slurs Every slur has a neutral counterpart that exists at a lower descriptive level in the sentence. It is called the Non-Pejorative Correlative (NPC) and substituting one term for the other is a truth-conditionally, but not expressively, equivalent replacement. Hom [3] formally defines the NPC for a slur as the expression that picks out the supposed extension of the epithet without derogating members of that extension:

- 1. that means literally nothing to a n-slur
- 2. that means literally nothing to an African-American

Community focused slur usage is separated from other sub-categories of slur usage because it is generalizable to all community members, rather than a finite set that is referenced by the speaker. Attribution is, however, taken to be community-focused when it is accompanied by negation. The following example illustrates why this is so:

- 1. lol don't be a f-slur
- 2. lol (do not you) be a f-slur
- 3. lol (do not you) be homosexual



Figure 4. Attribution is taken to be community-focused when it is accompanied by negation. The inclusion of *do not* redirects the derogatory force associated with *f-slur* onto *homosexual*, rather than *you*.

Stand Alone Slurs A slur is stand-alone when it is the only, single or repeated, word in a phrase:

- f-slur
- n-slur! n-slur n-slur n-slur
- *t-slur*;))!!!!

Sexualization A slur is sexualized when it is used as imagery for the purpose of fetishization, pornography, prostitution, or objectification. We isolate sexualized pejorative language from language in a violent or humorous context because the inclusion of the slur introduces a transactional element to derogation. Not only is the target community degraded, but, especially in the case of pornography, they are portrayed as a sexual object for consumption.

- A very hot black t-slur is mastu http://t.co/OfGGMLXQAj.
- I love the taste of a nice hot t-slur load
- it wasn't just hardcore porn, it was hardcore t-slur porn! fact check'd!

Self-Deprecation A slur is used for self-deprecation when it is negatively attributed to the speaker. Its appropriative counterpart is slur reclamation, which is characterized by a positive attribution to the speaker.

Appropriative Slurs

Appropriative slur usage is defined as meaningful usage by the targeted group for an alternate, non-derogatory purpose. It contains one subcategory:

1. Reclamation

Reclamation A slur is reclaimed when it loses its derogatory force, variation, and autonomy [3]. Applications of slur reclamation include, but are not limited to, self-, individual-, and group-targeted empowerment. These may take the form of positive attributions or connotations:

- i just don't see the point in pointing out how different we are when we're trying to be treated as equal. not all of it is about that but some of it is and it's that part that gets recognized by the public. i also feel no association to any of it. i'm a *f-slur* because i like dick...not rainbows. i hate rainbows.
- well now i kinda feel like the odd *t-slur* out lol(sorry bout the t word guys, all in good spirits.) i think i tried it because ive read about alot of other transmen who enjoyed it after they began transitioning

• 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 (NDNA) Slurs

NDNA slur usage is meaningful usage by targeted or non-targeted groups for an alternate non-derogatory, non-appropriative purpose [3]. A comment falling under this use case retains the derogatory force of the slur. Its subcategories are:

- 1. Referential Language
- 2. Counter Speech
- 3. Sarcasm

Referential Language

Discussion of the slur, its origin, or acceptable use cases;

- \bullet ask him if he would say n-slur if he had a black roommate
- I didn't really know t-slur was that inappropriate.
- I always get downvoted for saying t-slur or shemale
- Just 5-6 years ago you could say f-slur and noone would care

Direct Quotes, excluding paraphrasing and quotes involving non-embedded standalone slurs;

- what a about when harry truman said i think one man is just as good as another so long as he's not a *n-slur* or a chinaman?
- there is no dead *n-slur* storage. you know why? cause storing dead n-slur ain't my fuckin' business that's why!
- the day dave found out the girl he likes is a *t-slur* screenshot vote history on srscmharts < this comment posted by a bot | report an error
- actual quote: "de *n-slur* woman is de mule uh de world so fur as ah can see."

Recollection of a time when the slur was used;

- i used to troll this sub until i realized that it's literally a bunch of teenaged boys who are using what must be their newfound knowledge of sex and stds to see how many different ways they can call someone a *f-slur* so they can impress the older neighbor kids next door.
- one of my friends who happens to be a lesbian has always called me *t-slur*-butt as a term of endearment. to be honest, being called sir is far worse than being referred to as a *t-slur* for me...
- she says f-slur the whole time and you can tell she takes great joy in throwing it around
- i get called a f-slur by some random clitdick almost every day i have class.

Counter Speech Counter speech is a discouraging response to derogatory content. A slur used in counter speech responds directly to an instance of derogation, in defense against a comment made by a single speaker or group.

- Imao cmon fam we JUST talked about not using the word dyke.
- anyone can say nigga or *n-slur* as long as they don't use it in a derogatory sense.

Sarcasm A slur used for non-derogatory sarcastic purposes inverts the intuitive truth-conditions of a sentence [1]. As Wilson [6] argues, such is the case when speakers endorse an embedded, ironic meaning that is contrary to the one that they literally convey. For example in "imagine the difficulty of being white though, you have to cross the street every time you see a black man on the sidewalk, that could get exhausting, plus, you won't be able to say *n-slur* in public anymore, how are you supposed to sing along at the rap concert that you paid for?? think about it." the implicature that being white is difficult is inverted by sarcasm to draw attention to cultural appropriation and the derogatory *n-slur*.

Sarcasm is made explicit by Reddit users via the signals: s, rj, #sarcasm, or #s.

- 1 t-slur year = 5 cis years
- yes. straight people are welcome to apply for their *f-slur* license. we'll start you off with a *f-slur* learner's permit, which permits you to say the f-word between the hours of 8am and 11pm, only in the presence of a fully-licensed homosexual. after a trial period, you may be approved for wider use.

Homonyms

A homonym contains one or more dual, non-derogatory alternative meanings ¹.

- 1. *chink*/n. (i) a small cleft, slit, or fissure; (ii) a weak spot that may leave one vulnerable; (iii) informal, an ethnic slur usually referring to a person of Chinese ethnicity;
- 2. fag/v. (i) to tire by strenuous activity; n. (ii) cigarette; (iii) unit of measurement; (iv) used as an insulting and contemptuous term for a male homosexual;
- 3. t-slur/n. (i) vehicle transmission; (ii) slang for transgender person;
- 4. bitch/n. (i) the female of the dog or some other mammal; (ii) informal, a malicious, or overbearing woman; used as a generalized term of abuse and disparagement for a woman; (iii) informal, something that is difficult, objectionable, or unpleasant; (iv) informal, complaint;
- 5. kike/n. (i) insulting and contemptuous term for a Jewish person; (ii) nickname of Enrique;
- 6. dyke/n. (i) British spelling of dike; (ii) offensive, lesbian; (iii) reference to Dick van Dyke;

Examples are shown below:

- chink in the armor
- but as a whole we form a mighty f-slur
- i wouldn't buy a used honda *t-slur* from that era. that's when they had a spate of crappy transmissions and extended the warranties on them until over 100km

References

- [1] Elisabeth Camp. "Sarcasm, pretense, and the semantics/pragmatics distinction". In: Noûs 46.4 (2012), pp. 587–634.
- [2] Philip B Gove. "Noun Often Attributive" and "Adjective". In: American Speech 39.3 (1964), pp. 163–175.
- [3] Christopher Hom. "The semantics of racial epithets". In: The Journal of Philosophy 105.8 (2008), pp. 416–440.
- [4] Robert Truswell. "Attributive adjectives and the nominals they modify". PhD thesis. Citeseer, 2004.
- [5] Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. "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). 2019, pp. 602–608.
- [6] Deirdre Wilson. "The pragmatics of verbal irony: Echo or pretence?" In: Lingua 116.10 (2006), pp. 1722–1743.

¹Definitions by Merriam-Webster Dictionary