Alexis Palmer, Christine Carr, Melissa Robinson, Jordan Sanders

COLD: Annotation scheme and evaluation data set for complex offensive language in English

Abstract

This paper presents a new, extensible annotation scheme for offensive language data sets. The annotation scheme expands coverage beyond fairly straightforward cases of offensive language to address several cases of complex, implicit, and/or pragmatically-triggered offensive language. We apply the annotation scheme to create a new **Complex Offensive Language Data Set for English** (COLD-EN). The primary purpose of this data set is to diagnose how well systems for automatic detection of abusive language are able to classify three types of complex offensive language: reclaimed slurs, offensive utterances containing pejorative adjectival nominalizations (and no slur terms), and utterances conveying offense through linguistic distancing.

COLD offers a straightforward framework for error analysis. Our vision is that researchers will use this data set to diagnose the strengths and weaknesses of their offensive language detection systems. In this paper, we diagnose some strengths and weaknesses of a top-performing offensive language detection system by: a) using it to classify COLD, and b) investigating its performance on the 10 fine-grained categories supported by our annotation scheme. We evaluate the system's performance when trained on five different standard data sets for offensive language detection. Systems trained on different data sets have different strengths and weaknesses, with most performing poorly on the phenomena of reclaimed slurs and pejorative nominalizations. **NOTE:** *This paper contains sensitive and offensive material. The offensive materials are part of a complex puzzle we wish to better understand; they appear in the form of lightly-censored slurs and degrading insults. We do not condone this type of language, nor does it reflect the attitudes or beliefs of the authors.*

1 Introduction

Existing systems for detecting offensive language[1] mostly rely on identifying explicitly offensive keywords, and they often get things wrong, sometimes in rather troubling ways. Current evaluations tend to misrepresent the capability of these systems to handle the complex nature of offensive language (Wiegand et al., 2019, among others). Systems are particularly challenged when it comes to classification of implicitly offensive

[1] We use "offensive language" as a broad term to include hate speech, abusive language, insults, profanity, and other forms of expression triggering negative reactions in (at least some) hearers/readers.

language. At the same time, systems tend toward false positives on expressions like reclaimed slurs, with outcomes that unfairly and disproportionately penalize certain linguistic and social communities. Mis-classifications of such instances (false positives, in particular) contribute to the systematic racial bias seen in automatic offensive language classification (Davidson et al., 2019). That bias in turn can have a disproportionate negative effect on members of already-stigmatized communities.

In the name of better understanding existing systems' limitations, we present COLD (Complex Offensive Language Dataset),[2] an annotation scheme and evaluation data set that codes for reclaimed slurs as well as two categories of implicitly offensive language: distancing and pejorative adjectival nominalization. We propose the term ***complex offensive language*** to describe occurrences of offensive language signaled with more complex means than explicitly offensive keywords, as well as explicitly offensive keywords used with neutral or positive force. This term encompasses a broad range of linguistic phenomena using different means for conveying offense, or using purportedly offensive language in non-offensive contexts. COLD is easily extensible to include additional types of complex offensive language, and we hope that the community will join us in including data and annotations for additional types of complex offensive language.

Our vision is that this data set can be used to diagnose the strengths and weaknesses of offensive language detection systems, as well as offer a straightforward framework for error analysis. The motivation is similar to targeted evaluation data sets constructed in order to gauge the ability of automated systems to handle particular linguistic concepts. Our approach is also similar in spirit to efforts like the *Build it, Break it, The Language Edition* shared task (Ettinger et al., 2017) or the Winograd Schema Challenges (Kocijan et al., 2020).

Contributions. The contributions of this work are:

1. Linguistically-grounded discussion of several categories of complex offensive language;
2. A straightforward annotation scheme with simple annotation questions leading to fine-grained categorization of instances;
3. A procedure for attempting to mitigate the residual harms of annotation;
4. An evaluation data set with fine-grained categories relevant for detailed error analysis of systems for automatic detection of offensive language;[3] and
5. Diagnostic analysis of how well some existing systems are able to classify these three categories of complex offensive language.

Overview. Our approach to annotation targets three subtypes of offensive language that, we hypothesize, tend to elude correct classification by automatic systems. Each of

[2] In a previous presentation (Carr, Robinson, & Palmer, 2020), we used the name DEIOLD for the same data set.

[3] The data sets, related code, and annotation guidelines are available at: `https://github.com/alexispalmer/cold`.

these subtypes is connected to a significant linguistic literature, discussed in Section 2. The COLD annotation scheme (Section 3) poses four binary questions to annotators and then uses the answers to these questions to associate a single fine-grained category with each instance. We compile a diagnostic corpus from several different data sources (described in Section 4) and annotate it using the new scheme, creating a set of 2016 instances with both detailed linguistic annotation and fine-grained categorization.[4] We train neural models (Section 5) on five different data sets and analyze the models' performance on the new COLD corpus (results in Section 5.3).

2 Selected linguistic phenomena in complex offensive language

Offensive language that uses means other than explicitly offensive slur terms or profanity to convey offense is harder to automatically detect than explicitly offensive language. Such non-explicit phenomena are often referred to as *implicitly offensive language* (Wiegand et al., 2019; Waseem et al., 2017, among others). Another category that poses challenges for automatic detection systems are reclaimed or reappropriated uses of slur terms. Though identical in form with explicit slur terms, they serve entirely different semantic and pragmatic functions (Rahman, 2012). In this paper, we propose the term **complex offensive language** as an umbrella term for the range of linguistic phenomena posing challenges for automated detection of offensive language online.

The space of phenomena constituting the category of complex offensive language has yet to be fully delineated, and we suspect the members of this category are many and varied – offensive language is inherently a complex enterprise. The current work addresses three different phenomena, each of which we find to be both important and linguistically interesting.

The phenomena are realized in COLD as features of the annotation scheme. This section touches on some of the linguistic background for each of these features, each of which is complex and well-studied. We offer only a brief introduction to the literature and encourage the reader to consult cited works for more detailed and nuanced discussion.

First, we consider the phenomenon of **distancing** and several linguistic realizations of distancing (Sec. 2.1). Next, we consider **slurs** in both their typical uses and in the case of slurs reclaimed by the communities they were originally intended to degrade (Sec. 2.2). The third relevant characteristic is pejoration which arises through **adjectival nominalization**, described in Sec. 2.3.

2.1 Pejoration through distancing

Culpeper (1996) describes distancing/othering as a linguistic act through which a speaker creates space between themselves and another person or group. The distinction between **in-group** and **out-group** is crucial for understanding the use of distancing to cause offense. Following Staszak (2009), the in-group is a group the speaker belongs to,

[4] COLD has since been expanded to 2500 instances. Both the smaller and larger versions of the data set can be found in the repository.

and the out-group is a group the listener or some other individual belongs to. Speakers may use language to show that they think their own social group (in-group) is in some way superior to the out-group, or to show that the out-group has negative qualities. The resulting utterances are offensive but not easily detectable by automatic means, particularly when there is no offensive keyword to strengthen pejorative meaning.

Recent computational work has investigated the use of distancing-related features for hate speech classification (Alorainy et al., 2019; Burnap & Williams, 2016), but to our knowledge there is no previous data set which codes for the presence of distancing.

2.1.1 Impoliteness and distancing

The use of linguistic distancing to create offense is, in essence, a form of impoliteness. The study of impoliteness in language examines how the intent of a speaker can cause social disharmony, often by attacking the face of their interlocutor. Following Brown et al. (1987) and others, **face** can be viewed as public self-image, and conversational ideals should preserve face, thereby maintaining social harmony. Positive face represents a speaker's "desire to be liked and appreciated by others."

The use of offensive language creates disharmony in conversation (and in society), and Culpeper (1996) identifies various strategies speakers use to create disharmony; here we focus on the strategy of positive impoliteness. Positive impoliteness is an attack on the hearer's positive face, their desire to be accepted (Bousfield & Locher, 2008). Speakers may use derogatory words (1)[5,6] to target either an individual or a particular characteristic or characteristics seen as negative by the speaker. Speakers may also use taboo words (2) to cause a face attack, or demonstrative pronouns (3) to distance themselves from the targeted group.

(1) Ceasefire? Let's see how long those t*wel h**ds can go without trying to attack Israel then cry to national media when they get popped again ! [HateBase Twitter]

(2) Thinking that i care i don't give a f*ck about what you think ! [HateBase Twitter]

(3) When I talk about those blacks, I really wasn't talking about you. [HateBase Twitter]

Personal pronouns are another potential distancing strategy, as they can be chosen to represent in- and out-groups, emphasizing the polarization and distancing of the targeted group. Alorainy et al. (2019) confirm that pairs of pronouns are more frequent in hate speech than in neutral utterances. In particular, speakers may choose inclusive vs. exclusive pronouns (*we* vs. *they*, *us* vs. *them*, *ours* vs. *theirs*) to create a dichotomy in which the in-group is seen in a positive light and the out-group more negatively (Riggins, 1997). Examples (4) and (5) illustrate:

[5] We have made the choice to censor quoted slurs and profanity throughout the paper by masking initial vowels with asterisks. We also replace usernames with the token @USER.

[6] Throughout, the source of examples is indicated in square brackets. "HateBase Twitter" refers to the corpus presented in Davidson et al. (2017).

(4) Homelessness, is it our problem or someone else's? Granted the homeless are down on their luck and don't really have a choice weather or not they are poor, but that is not my fault. [(Pandey, 2004)]

(5) @USER all them b*tches far as f*ck from us. *We* got el gran p*llo up this way and it serves us well. @USER [HateBase Twitter]

The term **othering** is sometimes used to describe distancing constructions that use metaphor and stereotypes to highlight distance between two social groups, with the targeted group being negatively viewed by the speaker. One classic othering construction has the form 'X is `ADJ` for `NP`', as seen below. Such constructions are implicitly offensive when the `NP` contains a **neutral counterpart**, or non-pejorative correlate (Hom, 2008). This is essentially a non-pejorative alternative for a potential slur term. In (6), *gay* is the neutral counterpart.

(6) He looks straight for a gay man. [AdjNom: Tw.][7]

(7) You're pretty white for a Mexican, no that means you're prettier. [AdjNom: Tw.]

In example (6) the user creates opposition between *straight men* and *gay men* via the properties stereotypically associated with each group. Similarly, in example (7) the speaker creates distance between white people and Mexicans, reflecting stereotypes about skin color and beauty. Othering constructions can also occur with slurs, as in the back-handed compliment seen in example (8). The slur term makes the utterance explicitly offensive and therefore more easily detectable by automated systems.

(8) They are not like other *n*ggers*. [Twitter]

In this case, the speaker suggests that the grammatical subject of the utterance does not embody the negative properties stereotypically associated with the slur term.

2.1.2 Linguistic form, othering, and distancing

Linguistic form, for example the choice of a definite plural instead of a bare plural, also can be used as a means for the speaker to indicate non-membership in a referenced group. Acton (2014) argues that the definite plural indicates non-membership, as in example (9), in which the speaker suggests that they are not a member of the referenced group, *Americans*. This stands in contrast to the bare plural form in example (10), which makes no commitment as to the speaker's membership (or not) in the referenced group.

(9) The Americans love their fast food! [Acton 2014]

(10) Americans love their fast food! [Acton 2014]

Additionally, Acton points out that certain words, such as *gay*, can take on a derogatory meaning when used in the definite plural form. This is because *gay* has associated

[7] Twitter example from the AdjNom corpus (Robinson, 2018), described in Section 4.

social meaning from multiple statements of othering and marginalization. This social meaning adds distancing to non-membership, which can lead to pejorative meaning.

2.2 Slur terms: use and reappropriation

Slurs are lexical items that convey *negative attitudes* or *heavy emotional connotations* towards a social group (Hess, 2019), typically centered around race, nationality, religion, gender, or sexual orientation (Bianchi, 2014). As one of their functions, slurs derograte targeted groups or individuals, and they specifically call up one or more descriptive attributes of the targeted group.

Sometimes slurs are reclaimed by the community they are intended to oppress, through a process known variably as reclamation, (re-)appropriation, and resignification. When reappropriated, slurs shift in meaning, losing their pejorative load (usually, though not always). Reclaimed slurs are used in many different ways by different speakers in the community, often performing complex identity work. Rahman (2012) is an engaging and detailed look at African American communities and reclamation of the n-word. Theoretical approaches to understanding the variable meanings of slur words have developed from content-based analyses (Hornsby, 2001, for example), which struggle to capture the flexibility of slurs with respect to pejorative content, to more pragmatic accounts (Hom, 2008; Bianchi, 2014; McCready & Davis, 2017). The latter accounts have different views on the details of the interaction between contextualized usage and pejorative association, but share the understanding that appropriated in-group uses of slur terms occur in a context which alters the derogatory force of the word. Further, Hornsby, Hom, and Bianchi all convincingly argue that in-group appropriation of slurs pushes back against their derogatory force, by echoing and subverting offensive uses.

Our primary interest in reclaimed slurs for this work is the following: these uses, judged as non-offensive by human moderators, are very frequently incorrectly flagged as offensive or inappropriate by automated content moderation systems. The high frequency of these false positives leaves some speakers in the difficult situation of having to choose between a) being unfairly penalized for utterances that are unproblematic from their perspective (and, importantly, their community's perspective); or b) censoring their preferred use of such constructions for fear of triggering false flags.

Slurs as slurs. In (11), the Twitter user quoted attacks another user, using the term *f*ggot* to degrade the user and to indicate that the speaker sees gay men as inferior to other social groups. Bolinger (2017) proposes that all slurs have five characteristics in common: they are offensive even when the speaker does not intend them to be, they are offensive even to hearers who are not being targeted, they carry derogatory cultural meanings, they do not all appear to be equally offensive on the surface, and they can occur, sometimes, inoffensively (Bolinger, 2017).

(11) @USER fight me 1 on 1 ur choice of game f*ggot see what happens [HateBase Twitter]

(12) @USER not as sick as you m*zzie trash @USER [HateBase Twitter]

Slurs differ from other general pejorative terms (e.g. **sshole*, *sh*thead*). Pejorative lexical items generally target one or more specific individuals on a personal basis (13), and slurs target individuals based on characteristics of the group (Hay, 2013; Blakemore, 2015; Bach, 2018). The slur in (14) is a term that degrades based on characteristics of an entire religion; the degree of offense is much greater than in (13).

(13) No matter what there is always an *sshole wearing a Yankee cap at these games. [HateBase Twitter]

(14) @USER Send all the f*cking m*zzies home, now DOJ come and get me motherf*ckers. [HateBase Twitter]

Reclaimed slurs. In some cases, slurs may undergo a process of reclamation, or appropriation, through which people belonging to the targeted group use the slur in non-derogatory ways within the targeted community (Hess, 2019). As noted above, reclamation serves many different functions for the community, including (among others) expression of a sense of solidarity and companionship, signaling shared identity and socio-cultural experiences, performing the social and political move of "taking back" a word of violence and oppression, subverting social norms, and expressing friendship or other close relationships.

McCready and Davis (2017) offer what they call an invocational account of slurs, whereby the use of a slur term "invoke[s] a preexisting complex of social attitudes and background related to the slurred group." Under this account, the hearer's interpretation of the speaker's intention in using the slur depends crucially on whether the speaker and hearer each belong to the slurred group or the privileged (i.e. non-targeted) group. The speaker/hearer configuration determines whether the slur is used for subordination of the targeted group member, for expression of solidarity, for indicating complicity, or to make an accusation. For example, expression of solidarity may occur when the slur is uttered by one member of the slurred group to another member of the slurred group.

In (15), an African American Twitter user calls his friend *my n*gga* to signify that both belong to the group and, further, that they have a close relationship. The choice of this phrase over other available terms conveying shared identity (such as *brother*) signals that speaker and hearer are especially close (Rahman, 2012, 148).

(15) Imma have to f*ck my n*gga up with some pot brownies [HateBase Twitter]

In appropriation contexts, reclaimed slurs may be used by political organizations or artists to subvert socio-cultural norms (Hom, 2008). The text in example (16) is from a sign held by a woman at a rally for sex workers. Here the speaker subverts sexual norms placed on women by society by using a word that typically conveys negative attitudes about women.

(16) Sl*ts say yes [Text on a sign from an image on Twitter]

In rarer cases, terms that previously were highly-offensive slurs can, through repeated use and resignification, begin to be used in mainstream settings with neutral or even

positive meanings. One such example is *queer*. Though some speakers still use this as a slur (17), the reclaiming of the word has progressed far enough that positive in-community uses are not at all unusual, as in (18).

(17) Die f*cking qu**r. [HateBase Twitter]

(18) I'm literally laughing in shock, amazement, joy. In...just everything. In everything! Pop the champagne American qu**rs. We have arrived. [HateBase Twitter]

In some very specific contexts, which Bianchi (2014) describes as "highly-regulated," even speakers outside of the targeted community can use the expression non-offensively. For example, the phrases *Queer Theory* and *Queer Studies* are acceptable in academic settings (Bianchi, 2014), and *queer* is part of the widely-used acronym *LGBTQ(+)*.

2.3 Adjectival Nominalization

The pejorative use of adjectival nominalizations has become salient in public conversations in recent years. As an example, some public figures have come under fire for how they have referenced certain groups. In examples (19) and (20), the speaker uses *gay* and *black* in otherwise neutral contexts, and yet the particular linguistic form of these terms upset people (especially when spoken the way they were spoken).

(19) For the gays out there—ask the gays and ask the people—ask the gays what they think and what they do [Donald Trump]

(20) I have a great relationship with the blacks. I've always had a great relationship with the blacks. [Donald Trump]

Similarly, there was backlash on Twitter during the first GOP debate (August 6, 2015) over the use of the term *illegals*. Below are some examples of reactions to the multiple uses of *illegals* during the debate.

(21) The word "Illegals" will never cease to make me cringe y'all don't even pretend to see them as human beings #GOPDebate [AdjNom:Tw.]

(22) "Illegals" as a noun is SO PROBLEMATIC #GOPDebate [AdjNom:Tw.]

(23) You know when you are using an adjective as a noun you are being racist "felons" "illegals" "Blacks" #GOPDebate [AdjNom:Tw.]

In example (21), the speaker recognizes the dehumanizing effect of the nominalization, even if they do not pinpoint the linguistic source of the problem.[8] The speakers in (22) and (23) both remark that making the adjective *illegal* into a noun and referring to people in that way is offensive and/or racist. The examples above along with many other reactions to such linguistic forms show that people are aware of this subtle linguistic phenomenon, but the reasons behind it may not be widely understood.

Wierzbiecka (1986) explains that when adjectives are nominalized, the new nominal incorporates reference to a generic category or kind based on a prototypical idea of

[8] See Mendelsohn et al. (2020) for computational approaches to analyzing dehumanization in text.

that category. As seen in example (24) below, *blonde*, as an adjective, contains only the semantic adjectival property of hair color. However, as a noun, *a blonde* (25), it has new semantic properties that are associated with the prototypical idea of what a blonde (person) is. These properties include both factual information, such as +HUMAN, and stereotypical information, such as +DUMB and +SEX_OBJECT.

(24) ADJ: Becky has blonde hair. [Constructed]

(25) NOM: Becky is a blonde. [Constructed]

Some adjectives become strongly pejorative when nominalized. Robinson (2018) argues that the adjectives which undergo this transformation often have meanings tied to demographic features of individuals, such as socioeconomics, ethnicity, gender, or sexuality. When these adjectival demographic features are nominalized, the feature is expanded into a generic kind that is based on the prototypical idea of that class of individuals. This process leads to stereotypical properties, very often negative, being tied to the use of these adjectival nominalizations. Consider the contrast below:

(26) ADJ: Becky is a gay woman. [Constructed]

(27) NOM: Becky is a gay. [Constructed]

In example (26), *gay* is simply an adjective describing Becky's sexual orientation. Example (27), on the other hand, conveys negative sentiment toward Becky. Similarly, Dixon et al. (2018) observe the prevalence of identity terms in toxic comments and note that the syntactic frame in which the term appears can influence whether its use is neutral or offensive.

Robinson performs a focused, in-depth study of four adjectival nominalizations: *poor*, *gay*, *female*, and *illegal*. Additionally, she investigates variation between four different forms the nominalizations can take: indefinite singular (*a gay*), definite singular (*the gay*), bare plural (*gays*), and definite plural (*the gays*). Along with the nominalization process, changes to both reference type and linguistic form can influence the amount of pejorative meaning associated with an expression. Note the examples below.

(28) That moment when you realize ALL illegals are technically criminals. [AdjNom:Reddit]

(29) jesus christ i make a joke and now im a gay? [AdjNom:Tw.]

While the generic reference forms (bare plural and definite plural) can still be pejorative such as in example (28), the offensive meaning is enhanced when a specific individual is evoked using an adjectival nominalization, as in example (29). In addition, annotation studies (Palmer et al., 2017) confirm that nominalizations of these four terms are significantly more likely to convey offense than adjectival uses.

We investigate pejorative nominalizations in this work because they are frequent and insidious in their offensiveness. A straight keyword approach is certain to fail, given that these word forms are typically neutral when used adjectivally but offensive when used as nouns.

Summary. Our goal is to construct an evaluation data set that contains roughly equal numbers of offensive and non-offensive instances, with a reasonably balanced distribution across the categories of offensive slur, reclaimed slur, distancing (pejorative or not), and adjectival nominalization (pejorative or not). We select the latter three phenomena because they are difficult to detect automatically. The remainder of the paper describes the resulting data set and its use for targeted evaluation and error analysis.

3 COLD: The annotation scheme

The annotation scheme has two steps. For each instance, a) four Yes-No questions capture different characteristics of the instance; and b) a fine-grained category label is derived based on the answers to the four questions. The scheme can easily be extended to other types of complex offensive language, simply by adding questions and/or fine-grained categories as needed.

The four questions. The four questions to be answered for each instance are:

1. Is it **Offensive**?
2. Is there a **Slur**?
3. Is there an **Adjectival Nominalization**?
4. Is there **Distancing**?

For each question, the only possible answers are Yes and No. This streamlined process does not ask annotators to determine a final category label for a given utterance, only to determine whether the tweet is offensive overall and whether it contains any of the three linguistic phenomena. The approach of breaking a difficult annotation task down into a number of easier questions is inspired by Friedrich and Palmer (2014). Our approach also allows for easy annotation of instances which contain two or more different phenomena, such as a slur plus distancing.

Fine-grained categories. From the answers to the four questions, we deterministically derive a fine-grained category for the utterance (see Table 1). There are 10 possible categories, 5 of them offensive and 5 non-offensive.

In terms of categorization, we allow the presence of a slur in an utterance to take precedence over the other two linguistic features (nominalization and distancing). In other words, if an utterance has been labeled as containing a slur, only two of the ten categories are available. If the utterance is labeled offensive, it gets the category label *OffSlur*. If not, its category is *Reclaimed*. This decision is taken with error analysis in mind, so that utterances with and without slur keywords are in different groups with respect to error analysis. Other researchers using COLD could make a different decision in terms of the grouping of the instances into fine-grained categories.

Offensive utterances without slurs fall into one of four different categories: *OffNom* for offensive nominalizations; *OffDist* for offensive utterances with distancing; *OffBoth* for

Fine-grained category	Off?	Slur?	AdjNom?	Distancing?
Offensive with slur	y	y	y/n	y/n
Offensive with nominalization	y	n	y	n
Offensive with distancing	y	n	n	y
Offensive with both	y	n	y	y
Offensive, other	y	n	n	n
Reclaimed slurs	n	y	y/n	y/n
Nonoffensive with nominalization	n	n	y	n
Nonoffensive with distancing	n	n	n	y
Nonoffensive with both	n	n	y	y
Nonoffensive, no cues	n	n	n	n

Table 1: Fine-grained categories, defined with respect to four (yes-no) annotation questions.

offensive utterances with both nominalization and distancing; and *OffOther* for offensive utterances with none of the three cues. The categories are similar for non-offensive tweets without slurs: *NonNom*, *NonDist*, *NonBoth*, and *NonNone*.

Annotation guidelines and training. Each annotator attended two in-person training sessions with the authors. During the training sessions, important terms and concepts were discussed, including slurs, reclaimed slurs, adjectival nominalization, and distancing. Numerous examples were given to show what would and would not fall into each category. Additionally, the concepts of explicit and implicit offensive language were discussed, with slurs given as an example of explicit offensive language, and adjectival nominalization and distancing as examples of implicit offensive language. Annotators were given access to a set of written annotation guidelines which they could access online at any point in the annotation process.[9] The following are similar to the examples used for training: slurs (30), reclaimed slurs (31), distancing (32), and adjectival nominalizations (33).

(30) stfu! she thinks your ugly and hard to look at! get he f*ck out of chat f*ggot before i block you [HateBase Twitter]

(31) I love all my qu**r boys!! [HateBase Twitter]

(32) Michelle Obama is pretty for a black woman [HateBase Twitter]

(33) The uncultured poors complaining again? Lol [AdjNom: Tw.]

As part of each training session, each annotator completed a test batch with 50 instances, after which the annotators and the authors met to discuss results and disagreements. This is the only time that this sort of calibration and discussion took place. The first version of the annotation scheme used a different approach, asking annotators to select from a set of labels indicating different categories of offensive and non-offensive language. After the initial independently-submitted batches showed low agreement

[9] Our written guidelines are available in the COLD repository.

between annotators and some confusion about the labels, we switched to the 4-question procedure described above. To conclude training, each annotator independently labeled and submitted a final test batch of 25 instances. The elapsed time between training and the start of annotation was roughly one week for most annotators.

4 COLD: The data set

COLD is a collection of instances extracted from pre-existing offensive language data sets. We use the existing labels to guide sampling from the existing corpora, and each instance is then re-annotated using the scheme described above. To compile COLD, we use focused sampling (Wiegand et al., 2019) from pre-existing data sets to collect a balanced number of instances that potentially belong to the categories of offensive with slur, reclaimed (non-offensive) uses of slurs, adjectival nominalizations, and distancing. For the latter two categories in particular, automatic extraction from unrestricted data is difficult, so other strategies are needed. For nominalization, we select data from a hand-collected and annotated corpus. For distancing, we apply heuristic filters based on syntactic patterns. Focused sampling is appropriate precisely because the data set is intended for focused, diagnostic evaluation and error analysis. It is not intended for use as training data.

This section describes the filtering process over existing data sets (Sec. 4.1), the annotation process (Secs. 4.2 and 4.3), and the resulting final corpus (Sec. 4.4).

4.1 Focused filtering

We select 2400 tweets, comments, and posts, balanced across offensive and non-offensive instances, according to the original annotations in the corpora we sample from. The majority of instances (1860) come from the **HateBase Twitter** corpus of Davidson et al. (2017). This is supplemented by 140 instances from the **WaseemHovy** corpus (Waseem & Hovy, 2016), and 400 from Robinson's **AdjNom** corpus (Robinson, 2018).

Slurs and reclaimed slurs. All instances containing slurs or reclaimed slurs come from the HateBase Twitter corpus, which consists of 25,000 tweets annotated as offensive, hate speech, or non-offensive. As the name suggests, all data comes from Twitter and was extracted using the hate speech lexicon from HateBase[10] via the Twitter API. Original annotations were done via crowd sourcing, with each tweet receiving from 4-9 annotations. The resulting distribution is 5 percent hate speech, 76 percent offensive speech, and the rest neither hate speech nor offensive.

We extract instances containing slurs by querying the corpus for a set of 12 slur terms, then filtering for the number of annotators (minimum 3) who labeled the tweet as offensive and selecting the top 500 results. Reclaimed slurs are selected using a set of 10 frequently-reclaimed slur terms, then filtering for the number of annotators who

[10] `http://Hatebase.org`

labeled the tweet as non-offensive/neither, and taking the top 500 results. In the case of reclaimed slurs, most selected instances have 2-3 offensive labels and 4-6 non-offensive.

Distancing. To select tweets with potential **distancing**, we rely on heuristic patterns over part-of-speech labels; this particular filtering process is the least successful (see Sec. 4.5 for analysis), as these constructions are the most difficult to detect automatically.

We first tag the HateBase Twitter corpus with part-of-speech labels, using the Stanford POS tagger (Toutanova, Klein, Manning, & Singer, 2003). We then search for particular POS patterns, to retrieve othering expressions (of the form '`ADJ PREP NP`') and distancing using personal and/or possessive pronouns (particularly tweets which set up us/them dichotomies). The retrieved expressions are manually inspected, and for othering expressions, we enforce the constraint that the expression includes a non-pejorative correlate (see also section 2.1). From the HateBase corpus, we extract 100 tweets with othering constructions and 260 with distancing using pronouns.

To reach the target of 500 tweets with potential distancing constructions, we sample additional data from the corpus created by Waseem and Hovy (2016). This corpus consists of about 16,000 tweets, annotated as racist, sexist, or neither. We use the same sampling process applied to the HateBase Twitter corpus. As a result, we extract 140 instances of distancing using pronouns.

Adjectival nominalizations. Finally, we include 400 instances of adjectival nominalizations, selected from the data from Robinson (2018). The corpus includes both nominalizations and adjectival uses of four target forms: *poor*, *female*, *gay*, and *illegal*. The nominalizations are distributed across four linguistic forms: indefinite singular, definite singular, bare plural, and definite plural. Each instance is annotated for both pejoration and linguistic form.

The corpus, which was collected for linguistic research on pejorative nominalizations, contains 1742 nominal forms across the 4 target words, plus adjectival uses. Each instance was manually identified and verified, for example, ensuring that occurrences of *female* have human reference. The instances come from a variety of sources, including reddit, Twitter, YouTube videos and commentary, blogs, and news media. Searches were conducted across various platforms on topics likely to produce high levels of the target forms. For example, subreddits such as The_Donald have threads on the topic of immigration. Most of the data was collected during the 2016 US Presidential elections, and many of the topics center around that election and issues central to the election, such as immigration. Many of the blogs, videos, and subreddits searched cater to the Alt Right or to Men's Rights groups such as Men Going Their Own Way (MGTOW). Other data comes from more neutral sources, such as a CNN interview with a business woman sharing her political beliefs about the potential of a female president.

Search engines work well for finding plural forms. Singular forms, however, present more of a challenge, as search results often return adjectival forms rather than singular noun forms. To resolve this issue, common verbs, such as forms of copular *be*, were added after the singular form while searching, resulting in instances like example (34).

(34) Yeah dude being poor happens from time to time, but being A poor is a way of life. LOL [AdjNom:Tw.]

From the adjectival nominalization corpus, we randomly select 300 instances annotated as pejorative and 100 instances annotated as nonpejorative. These are divided evenly across the four target forms, with 75 pejorative and 25 non-pejorative instances for each lexical item.

4.2 Annotation Process

We trained and employed six undergraduate Linguistics majors for this annotation task. With the goal of 2500 annotated instances, each with labels from three different annotators, we assigned each annotator a total of 1250 instances, divided into batches of either 100 or 50 instances each. Annotators were paid $12/hour for their work, and the batches were distributed through online spreadsheets, allowing annotators to work from their preferred locations. Each batch was assigned to three annotators, shuffling the groupings of annotators from batch to batch. For example, batch 1 was assigned to annotators A-B-C, batch 2 to D-E-F, batch 3 to A-B-D, batch 4 to A-C-E, and so forth.

The corpus is compiled from several pre-existing data sets (see Section 4). We distribute instances from the two smaller data sets evenly across the batches.

Mitigation of the residual harms of annotation.[11] Given the highly toxic nature of the data, and the growing evidence suggesting that regular and repeated exposure to such data can be detrimental (Newton, 2019; Simon & Bowman, 2019; Dwoskin, Whalen, & Cabato, 2019), we took measures to mitigate potential negative effects. First, we openly discussed the potential issues with our annotators, encouraging them to work on the data in small chunks of time, to perform self check-ins of their mental states, and to take breaks as necessary. We checked in with them periodically and encouraged them to reach out to us with any concerns. Second, we incorporated hand-selected cute animal videos into the annotation batches, inserting a video link after every 25 instances. The videos were offered as a required step in the annotation process, with link text reminding the annotators (e.g.) "Watch this cute video for a mental health break. It's required, and you get paid for doing it!" Anecdotally, the annotators have said that they enjoyed watching the videos and that they did provide a welcome break from the difficult data. Our understanding of the effectiveness of this strategy is limited. A careful and rigorous study of the potential harms of annotating toxic data is needed in the future, as well as studying the effectiveness of various strategies for mitigating those harms.

The potential negative impact of extended exposure to toxic data is not to be underestimated, and we feel that some sort of mitigation strategy should be incorporated whenever annotators are asked to take on this kind of task.

[11] Thanks to an anonymous reviewer for the phrase "residual harms of annotation."

	COLD: 2016 instances	
	Majority Y	**Majority N**
Offensive	952	1064
Contains Slur	1012	1004
Adj. Nom.	500	1516
Distancing	89	1927

Table 2: Distribution of labels in the COLD corpus.

4.3 Inter-annotator agreement and selection of gold-standard labels

We compute agreement between annotators for each of the four annotation questions using the `nltk.metrics.agreement` implementation of Fleiss' *kappa*, as described in (Artstein & Poesio, 2008).

Over the 2016 instances of the COLD corpus, we see good agreement for the categories of Offensiveness (0.61) and Distancing (0.76), and only moderate agreement for the categories of Slur (0.38) and Adjectival Nominalization (0.43). The label distribution for the four categories appears in Table 2. The high agreement seen for Distancing is an artifact of the skewed distribution for this label, as annotators marked very few instances of distancing.

The relatively low agreement we see for Slurs reflects the fact that this is quite a subjective assessment. We chose not to use any sort of lexicon of slurs, leaving annotators to decide for themselves whether an instance contains a slur. Although there is evidence to suggest that demographic factors influence human classification of offensive language (Waseem, 2016), we have not done a detailed study of this interaction. With the exception of race, the demographic profiles of our annotators are somewhat similar. All identify as female, and all are between the ages of 19 and 25. We did not investigate other demographic factors.

Additional training could be helpful for recognizing nominalizations as well, as this distinction relies on linguistic awareness of part-of-speech categories. Overall, inter-annotator agreement for this data set is moderate.

Gold-standard labels. To produce a stable labeled version of the data set, we take a simple majority vote for each annotation question, for each instance. Table 2 summarizes the distribution of labels in the final corpus, across the four annotation questions. Of the 2016 instances in the data set, 47% are labeled offensive, 50% as containing a slur, 25% as containing adjectival nominalizations, and only 4% as containing distancing.

Annotator self-consistency. Four months after the original annotations were finished, we asked three of the original annotators to re-annotate two batches of 50 instances each, in order to measure annotator consistency and reliability for the task. We selected one batch from early in the original annotation process (first 10 batches) and one from later

in the process (last 10 batches), selecting batches that were originally labeled by the three annotators in the consistency study. The annotators did not have access to their original data files. Table 3 shows the percentage of instances for which each annotator agreed with their original annotations, four months later. We see high agreement across the board, with the lowest agreement for the question of whether a tweet is offensive.

Annotator	Offensive?	Slur?	AdjNom?	Distancing?
Annotator A	87%	96%	92%	92%
Annotator D	87%	96%	96%	92%
Annotator E	94%	96%	99%	97%

Table 3: Consistency check – percentage of instances (out of 100) for which the annotator gave the same label as their original annotation, 4 months later.

4.4 Data set details and clean-up

As released, COLD consists of 2016 triply-annotated instances, meaning we lost nearly 400 of the extracted instances due to various problems with annotations. First, only 2156 of the instances were sent out to the annotators. After collecting, collating, and cleaning up the annotations, we found that the number of annotators per instance varied. 121 instances had labels from only 1 or 2 annotators and were therefore excluded from COLD. 206 instances had labels from 4 or more annotators. For each of these instances, we randomly selected 3 annotators and set the remaining labels aside. Finally, 19 instances were removed due to problems with ID numbers.[12]

After converting the Y/N labels to fine-grained categories (Sec. 3), we see the distribution of categories shown in Table 4. The corpus shows good representation for the categories of offensive with slur, reclaimed slur, offensive with nominalization, non-offensive with nominalization, and non-offensive (no cues). All categories related

[12] The publicly-available version of the data set includes all annotations. A second, slightly-larger version of COLD is also available, with 2500 instances.

Offensive types	#	Non-offensive types	#
Offensive with slur	620	Reclaimed slurs	392
Offensive with nominalization	201	Non-off. with nominalization	141
Offensive with distancing	19	Non-off. with distancing	6
Offensive with both	31	Non-off. with both	0
Offensive, no cues	81	Non-off., no cues	525
Total Offensive	952	Total Non-offensive	1064

Table 4: DEIOLD: number of instances per fine-grained category, based on a majority vote of three annotators. (Category definitions in Table 1.)

to distancing are under-represented, in some cases severely. For example, though we extracted 500 instances potentially containing pejorative distancing, annotators only mark 50 of them as offensive.

Clean-up of annotations. Following annotation, the following cleaning steps were taken to prepare the data for analysis. First, labels were modified as necessary for consistency (e.g. removing leading or trailing spaces, changing case). Next, the format of instance IDs was normalized to contain both a data set code and a unique ID number. New line characters in the middle of tweets were replaced with spaces, and duplicate annotations (i.e. same tweet, same annotator) were removed.

4.5 Problems with distancing.

There is a huge discrepancy between the number of instances we extracted as potentially containing distancing constructions ($n = 500$) and the number labeled by the annotators as containing distancing ($n = 89$). This discrepancy could be caused by several different factors: a) a noisy extracting/filtering process; b) annotator misunderstanding and/or confusion; c) poor training of the annotators; or d) faulty conception of the category.

One way of testing some of these factors is to have the same data labeled by the experts who conceived of the category. We perform a small study in which two of the co-authors labeled 98 instances previously extracted as potential distancing constructions. The experts used the four-question annotation process.

Table 5 shows the results. Of the 98 instances, the two experts agree on the distancing label for 56 instances, slightly more than 50%. 18 of the agreed-upon instances are marked by both annotators as *not* containing distancing. The examples below show the four possible label configurations: Yes-Yes, No-No, Yes-No, and No-Yes.

(35) **Agree: Yes** – Hearing these girls in my class talk about cars is some of the lowest IQ discussion I've ever heard. #NotSexist #JustAnObservation [WaseemHovy]

(36) **Agree: No** – Time to see these girls finally cook.... Sink or swim. Sassy won't save you #mkr [WaseemHovy]

(37) **Disagree: Yes, No** – 61% of welfare/government aid is claimed by white people. So y'all black slander is trash now. [HateBase Twitter]

	Offensive?	**Slur?**	**AdjNom?**	**Distancing?**
Agreement	84%	89%	99%	57%
	Distancing →	**AgreeYes**	**AgreeNo**	**Disagree**
Number of Instances		38	18	42

Table 5: Top: Agreement between two experts for 98 instances potentially containing distancing. Bottom: Number of instances per agreement status.

(38) **Disagree: No, Yes** – These girls should know skinny sausages are no fun at all. #mkr [WaseemHovy]

Where the experts disagree, they are remarkably consistent in their disagreements; in 32 of the 42 instances, Expert 1 chooses Yes and Expert 2 chooses No. This finding suggests that each annotator has a coherent notion of what is intended by distancing, but that there are some differences in their respective ideas of the category. Both experts participated in the training process, potentially leading to mixed messages about distancing coming through to the annotators.

Overall, at least one expert chooses the Yes label for about 80% of this randomly-selected subset of potential distancing constructions, but the two expert annotators only agree on distancing for 56/98. This result clearly indicates significant problems with both the extraction process and our understanding of what counts as distancing.

5 Using COLD for diagnostic evaluation

In the remainder of this paper, we put COLD to its intended use: as a diagnostic evaluation data set, to better understand the performance of offensive language detection systems on difficult categories of complex offensive language.

We use one model architecture – a pre-trained BERT (Devlin, Chang, Lee, & Toutanova, 2018) model fine-tuned on an offensive language data set – and train five different versions, each one fine-tuned on a different data set. There is no overlap between the data sets used for fine-tuning the BERT model and those from which COLD is compiled.

This section describes details of the model (Sec. 5.1) and data sets (Sec. 5.2) and results of the diagnostic evaluation (Sec. 5.3).

5.1 Model details

Due to its state-of-the-art performance on many NLP tasks, as well as its success in the Semeval 2019 offensive language detection task (Zampieri et al., 2019b), we use BERT (Devlin et al., 2018) as our base model. The top performing system in the Semeval task fine-tunes a BERT model on the task's training data set (Liu, Li, & Zou, 2019); we take a similar approach. Using this general architecture, we use five different offensive language and hate speech data sets with varying levels of label granularity, and we report performance for each of the five resulting models.

BERT. Bidirectional Encoder Representations for Transformers (BERT) was released by Google Research (Devlin et al., 2018) and has achieved state-of-the-art results on a variety of NLP tasks. The model is based on Vaswani et al. (2017)'s multi-headed transformer model and utilizes a masked language modeling and next sentence training regime. Typical language modeling schemes have the model attempt to predict the next token given a sequence of input tokens and repeat this process over the entire training set, requiring significant training time and compute resources. Masked language modeling,

on the other hand, masks a fraction of the input tokens and trains the model on only those that are masked, reducing training time as well as preventing the bidirectional model from leaking information about the token being predicted. In combination with masked language modeling, input sequences are also paired with true and false following sentences so the model is also forced to determine if it is the true next sentence. Finally, all outputs are concatenated with a classification token, so the model can be used for sentence classification (e.g. as offensive or not).

Several pre-trained base BERT models are available. These models for general English can be fine-tuned on task-specific data sets at a substantially-reduced time and compute cost compared to training a new model from scratch. The fine-tuning process performs the masked language modeling step and uses the target label provided by the data set as the concatenated classification token.

Data preprocessing. We follow the preprocessing techniques used by the winning team NULI (Liu et al., 2019) in the SemEval-2019 Task on offensive language detection in social media posts.

1. **Lower case all text** as the model we use is uncased.
2. **Replace urls with "http"**, preventing wordpiece from segmenting URLs into tokens that have likely never been seen in the training data.
3. **Limit consecutive @USER mentions to 3** to reduce redundancy, as users are mapped to a single placeholder token (@USER).
4. **Segment hashtags into component words.** Tweets often utilize hashtags with important keywords strung together, we use the python library `wordsegment` to separate these into their component tokens.
5. **Replace emojis with word descriptions.** Unicode characters for emojis often have poor embedding representations relative to the tokens describing them, so each emoji is converted to its textual description. (😊 → smiley face)

5.2 Fine-tuning data sets.

We perform the same fine-tuning process as Liu et al. (2019) on five different offensive language classification data sets. We use BERT's built-in tokenizer wordpiece, which splits words into their morphological components, and start with the BERT-Large, Uncased (Original) model with 12 transformer blocks, 12 attention heads, and 110 million parameters. We use the default fine-tuning configuration and train for 3 epochs. Of the five data sets, four use binary labels (Off. or Not), and one uses three different labels for offensive utterances. In the case of OLID-2020, the official shared task data is labeled in a semi-supervised fashion with a confidence score, and we follow Fromknecht and Palmer (2020) in choosing a confidence threshold to convert scores into binary labels. The data sets are summarized in Table 6.

Data set	Offensive	Non-Offensive	Total
HASOC_multi	HATE / OFFN / PRFN - 1,143 / 667 / 451	NOT - 3,591	5,852
HASOC_binary	HOF - 2,261	NOT - 3,591	5,852
OLID-2019	OFF - 4,400	NOT - 8,840	13,240
OLID-2020	OFF - 1,426,195	NOT - 5,834,139	7,260,334
Kaggle-Toxic	TOX - 16,225	NOT - 143,346	159,554

Table 6: Label distributions and statistics for fine-tuning data sets. (LABEL - #)

HASOC. The Hate Speech and Offensive Content Identification in Indo-European Languages data set (HASOC) (Mandl et al., 2019) consists of several thousand social media posts from Twitter and Facebook in English, Hindi, and German. Posts are labelled for three separate sub-tasks: (1) binary coarse grained labels hate/offensive (HOF) or NOT, (2) multi-class fine-grained labels (NOT/HATE/OFFN/PRFN), and (3) posts labeled HOF in task 1 are designated as targeted or untargeted hate/offensive language. The labels:

- Hate Speech (HATE) : Text applying negative attributes to an individual due to membership in a group or hateful comments towards groups because of race, politics, sexual orientation, gender, social status, health condition, or similar.
- Offensive Language (OFFN): Posts containing threats of violence or attempts to degrade, dehumanize, or insult an individual.
- Profanity (PRFN) : Language typically deemed inappropriate such as cursing or swearing that is not abusive or insulting in nature.

We fine-tune two BERT models from this data set using English training instances from tasks 1 (HASOC_Binary) and 2 (HASOC_Multi).

OLID-2019 and OLID-2020. The two OLIDs (Offensive Language Identification Dataset) come from two consecutive years of SemEval shared tasks addressing identification and classification of offensive language.

OLID-2019 (Zampieri et al., 2019a) contains thousands of tweets identified using keywords and labelled with a three layer hierarchical annotation scheme for offensive language. The first layer categorizes text as either offensive or not, the second layer specifies whether the offensive language is targeted or not, and the final layer categorizes the target as an individual, group, or other. Language is considered offensive in this data set if it contains insults, threats, or profanity. We use only the first layer annotations to fine-tune a model as a binary classification task.

OLID-2020 contains millions of tweets following the same three layer hierarchical annotation scheme. Unlike OLID-2019, there are no manually-provided gold-standard labels for OLID-2020. Instead, each tweet is labeled with confidence scores (ranging from $0.0 - 1.0$) derived by averaging over the output of a number of supervised systems

for detection of offensive language. Task participants are responsible for converting confidence scores to labels. Following Fromknecht and Palmer (2020), we use a threshold of 0.44 for the first annotation layer (offensive or not).

Kaggle-Toxic. The Kaggle-Toxic[13] data set is a collection of Wikipedia Talk page edits shared on Kaggle as a toxic comment classification competition. The texts are posts and replies by Wikipedia contributors discussing page edits that sometimes lead to arguments and toxic behavior. The data set is annotated for six labels that have some overlap, and often a post contains multiple labels. The labels are toxic, severe-toxic, obscene, threat, insult, and identity-hate. The two toxic labels are somewhat ambiguous, and no annotation guidelines were released for this data set. We remove the ambiguous overlap between classes and reduce the problem from a multi-label to a binary classification task of toxic/not. Any text annotated to have any of the original six categories is labeled as TOX, and any text containing none of the original categories is labelled as NOT.

In-domain classification accuracy. Classification accuracy for each model is evaluated on a development set (20% stratified split of the training data used for fine-tuning). On the original labels, development set accuracies are as follows: (a) HASOC_multi: 65.4%; (b) HASOC_binary: 71.2%; (c) OLID-2019: 79.2%; (d) OLID-2020: 97.5% ; and (e) Kaggle-Toxic: 96.2%. Kaggle-Toxic is an outlier due to both the larger amount of data and its skewed distribution, and the high performance for OLID-2020 reflects the much larger amount of training data available. It's worth noting that accuracy on the test data is significantly lower.

Evaluating these models on COLD means not only moving to a new domain for testing, but also testing on mixed-domain data. Thus we can reasonably expect lower overall performance than what we see on an in-domain development set. This cross-domain evaluation setting provides a realistic view of the performance to be expected when applying a system to new data.

5.3 Diagnosis of performance on fine-grained categories

The ten fine-grained categories associated with the instances in COLD allow us to investigate what kinds of labeling decisions each model makes for each type of instance. To use COLD in its intended diagnostic capacity, we compare the distribution of predicted labels for each category with the expected label, given the majority-vote annotations. It is important to note that the set of output labels used by any given model is determined by the labels in the data set used for fine-tuning the model. The OLID-2019 model, for example, outputs the labels OFF and NOT, where the HASOC_binary model outputs HOF and NOT. The output labels of binary models can be mapped straightforwardly to the distinction between offensive and non-offensive

[13] https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge

Type		HasocBinary OFF/NOT	OLID-19 OFF/NOT	OLID-20 OFF/NOT	KagTox TOX/NOT
OffSlur	620	77% / 23%	85% / 15%	87% / 13%	88% / 12%
OffNom	201	40% / **60%**	**64%** / 36%	**59%** / 41%	40% / **60%**
OffDist	19	68% / 32%	68% / 32%	**79%** / 21%	63%/ 37%
OffBoth	31	45% / 55%	**74%** / 26%	**71%** / 29%	23% / 77%
OffOth	81	74% / 26%	79% / 21%	80% / 20%	77% / 23%
ReclSlur	392	**79%** / 21%	**79%** / 21%	**83%** / 17%	**86%** / 14%
NonNom	141	34% / 66%	43% / 57%	35% / 65%	28% / 72%
NonDist	6	33% / 67%	77% / 23%	67% / 33%	50% / 50%
NonBoth	0	–	–	–	–
NonNone	525	22% / 78%	27% / 73%	28% / 72%	20% / 80%

Table 7: Diagnostic evaluation, training on different data sets and making predictions for COLD. The table shows the percentage of instances in each category assigned each model output label, for 4 binary models.

fine-grained categories. Our current approach to diagnostic evaluation involves looking at the percentage of instances within one COLD category assigned each of the training labels used by the model.

Results for the four binary models appear in Table 7, and results for the one multi-class model appear in Table 8, in both cases with interesting numbers in boldface. For example, the model fine-tuned on OLID-19 predicts OFF for 79% of the 392 instances in COLD's **ReclSlur** category, providing empirical verification that this model is terrible at accurately labeling reclaimed slurs.

Finding One: All models do a good job of handling the explicit categories , labeling offensive utterances with slurs as offensive in 72-88% of cases. Similarly, non-offensive utterances with no pejoration cues receive non-offensive labels 73-88% of the time, depending on the model. These results are expected; slurs serve as effective keyword features for all models, and the non-offensive cases without pejoration cues rarely lead the model astray.

Finding Two: All models fail at recognizing reclaimed uses of slur terms as non-offensive ; these are classified as offensive or toxic 69-86% of the time. For example, (39) is labeled by annotators as N-Y-N-N, yielding the label **reclaimed**, yet all four models classify the tweet as offensive or toxic. Interestingly, the multi-class model labels most of these as Profanity rather than Offensive.

(39) so many pretty b*tches coming to my birthday dinner [HateBase Twitter]

Type		HasocMulti HATE/OFF/PROF // NOT
OffSlur	620	05% / 21% / **46%** // 28%
OffNom	201	13% / 10% / 12% // **65%**
OffDist	19	32% / 26% / 10% // 32%
OffBoth	31	16% / 13% / 10% // 61%
OffOth	81	15% / 28% / 15% // **42%**
ReclSlur	392	01% / 04% / **65%** // 31%
NonNom	141	06% / 06% / 07% // 81%
NonDist	6	17% / 00% / 00% // 83%
NonBoth	0	–
NonNone	525	04% / 03% / 05% // 88%

Table 8: Diagnostic evaluation, training on different data sets and making predictions for COLD. The table shows the percentage of instances in each category assigned each model output label, for 1 multi-class model.

Finding Three: Three of four models tend to miss the pejorative meaning associated with adjectival nominalizations , mislabeling them about 60% of the time. OLID-19 and OLID-20, the exceptions, still get about 40% of such cases wrong. The utterance in (40) is clearly anti-gay and is labeled by annotators as Y-N-Y-N, yielding the label **OffNom**. All four models classify the utterance as non-offensive. In other cases, such as (41), non-pejorative nominalizations are considered offensive by most models.

(40) Another huge reason I'm against gays is because their role in politics. [AdjNom]

(41) As a gay teen I can confirm first hand that there are many gays that are depressed. [AdjNom]

The small number of occurrences of distancing identified by annotators prevent us from drawing any strong claims, but we can look at a representative example. (42) is a canonical pejorative othering construction, labeled by annotators as Y-N-N-Y. Only one model (Kaggle-Toxic) classifies it as toxic, and the other three mark it as non-offensive.

(42) You're pretty for a black girl [HateBase Twitter]

5.4 Discussion

Overall, the patterns seen in the analysis are not unexpected, but the structure provided by COLD allows us to easily see which categories are most difficult for current models.

Diagnosis using COLD reveals that, despite strong classification accuracy overall, state-of-the-art models fail hard on implicit offensive language, as well as on reclaimed slurs. The result on reclaimed slurs is particularly informative, as mis-classifications of such instances (false positives, in particular) contribute to the systematic racial bias

seen in automatic offensive language classification (Davidson et al., 2019). That bias in turn can have a disproportionate negative effect on members of already-stigmatized communities.

The research community has long been aware of difficulties with detecting implicit and complex offensive language. The new contribution made here is a mechanism for systematic investigation of the strengths and weaknesses of systems for automatically detecting offensive language. Such systematic investigation, coupled with extensive error analysis, holds promise for making dramatic improvements to models for automatic detection of abusive language.

The results are disappointing for the phenomenon of pejoration through distancing, though the source of the problem likely lies in data extraction and filtering, as well as annotation guidelines and training, rather than in the classification process, given that the corpus contains few instances labeled as containing distancing. Even expert annotators found that distancing occurs in fewer than 50% of the instances extracted as potential cases of distancing.

It is our hope that improving performance on these categories of complex offensive language will improve performance for offensive language detection overall. We also hope to identify and add new linguistic phenomena to the data set, perhaps starting our research with the instances in the **Offensive-Other** category.

We find cross-domain classification, as recommended by Wiegand et al. (2019), to be an appropriate setting for performing this analysis, and we hope that domain effects are mitigated at least to some extent by having assembled the corpus from disparate sources. As a next step, we would like to train additional BERT models, this time fine-tuning on HateBase Twitter and the Waseem and Hovy corpus (removing the instances already in COLD). This will show us whether the error patterns change when the model is trained and evaluated within the same domain.

6 Conclusion

This paper presents a new annotation scheme and evaluation data set for offensive language detection in English which incorporates awareness of reclaimed slurs, pejorative adjectival nominalizations, and pejoration through linguistic distancing. In order to achieve better handling of complex offensive language, we argue that these phenomena need to be considered. A system that can make correct predictions for these patterns will be a stronger system overall, as the phenomena included in COLD require more nuanced awareness of the roles of linguistic form, politeness strategies, and sociolinguistic factors. Rather than building a large training corpus, we start by developing an evaluation data set to diagnose whether these patterns are in fact mislabeled by existing systems.

We evaluate the performance of models trained on five different offensive language data sets and find some interesting patterns of correspondence between predicted labels and fine-grained categories. In particular, we show that current models perform poorly on reclaimed slurs and pejorative nominalizations, no matter which data set is used for model fine-tuning.

Looking ahead, we are considering several different directions for extending this work. First, the data has been made publicly available, together with annotation guidelines and model outputs for analysis.[14] We hope to continue expanding the corpus, both with additional data sampled in a less-restrictive way, and with focused data targeting specific linguistic phenomena. We plan to explore the possibility of collaborative data set expansion, deciding on a common format to make it easy for other researchers to contribute to COLD. In addition, we are considering developing a system to generate diagnostic reports for existing systems. We can use the multiple annotations to identify high-agreement (i.e. "easy") and low-agreement (i.e. "difficult") cases, enhancing the quantitative results with examples of errors the system makes across various categories.

Acknowledgments

First and foremost, our thanks to the three anonymous reviewers whose extensive feedback certainly made this article better. Next, thanks to our fantastic annotators: Tiffany Blanchet, Samantha Boyer, Kathryn Denier, Katie Eilerman, Meg Fletcher, and Meghan Renshaw. We have also benefited from helpful discussions with colleagues Patricia Cukor-Avila and Xian Zhang, as well as with students in LING 5550 at the University of North Texas. This work was made possible by a research seed grant from the Dean's Office of the College of Information at UNT.

References

Acton, E. (2014). *Pragmatics and the social meaning of determiners* (Doctoral dissertation, Stanford, CA). Retrieved from `http://www.emich.edu/english/faculty/documents/suthesisacton.pdf`

Alorainy, W., Burnap, P., Liu, H., & Williams, M. L. (2019). "The Enemy Among Us": Detecting Cyber Hate Speech with Threats-Based Othering Language Embeddings. *ACM Trans. Web*, *13*(3).

Artstein, R., & Poesio, M. (2008). Survey Article: Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, *34*(4), 555–596.

Bach, K. (2018). Loaded words: On the semantics and pragmatics of slurs. *Bad Words: Philosophical Perspectives on Slurs*, 60–76.

Bianchi, C. (2014). Slurs and appropriation: An echoic account. *Journal of Pragmatics*, *66*, 35–44.

Blakemore, D. (2015). Slurs and expletives: A case against a general account of expressive meaning. *Language Sciences*, *52*, 22–35.

Bolinger, R. J. (2017). The pragmatics of slurs. *Noûs*, *51*(3), 439–462.

Bousfield, D., & Locher, M. A. (2008). *Impoliteness in language: Studies on its interplay with power in theory and practice* (Vol. 21). Walter de Gruyter.

[14] `https://github.com/alexispalmer/cold`

Brown, P., Levinson, S. C., & Levinson, S. C. (1987). *Politeness: Some universals in language usage* (Vol. 4). Cambridge University Press.

Burnap, P., & Williams, M. L. (2016). Us and them: identifying cyber hate on Twitter across multiple protected characteristics. *EPJ Data science*, *5*(1), 11.

Carr, C., Robinson, M., & Palmer, A. (2020). *Improving hate speech detection precision through an impoliteness annotation scheme.* Retrieved from `https://www.linguisticsociety.org/abstract/improving-hate-speech-detection-precision-through-impoliteness-annotation-scheme` (Presentation at Annual Meeting of the Linguistic Society of America)

Culpeper, J. (1996). Towards an anatomy of impoliteness. *Journal of Pragmatics*, *25*(3), 349–367.

Davidson, T., Bhattacharya, D., & Weber, I. (2019). Racial Bias in Hate Speech and Abusive Language Detection Datasets. In *Proceedings of the Third Workshop on Abusive Language Online* (pp. 25–35). Florence, Italy: Association for Computational Linguistics.

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In *Eleventh International AAAI Conference on Web and Social Media.*

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.

Dixon, L., Li, J., Sorensen, J., Thain, N., & Vasserman, L. (2018). Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 67–73).

Dwoskin, E., Whalen, J., & Cabato, R. (2019). *Content moderators at YouTube, Facebook and Twitter see the worst of the web - and suffer silently.* `https://www.washingtonpost.com/technology/2019/07/25/social-media-companies-are-outsourcing-their-dirty-work-philippines-generation-workers-is-paying-price/`. The Washington Post.

Ettinger, A., Rao, S., Daumé III, H., & Bender, E. M. (2017). Towards Linguistically Generalizable NLP Systems: A Workshop and Shared Task. In *Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems* (pp. 1–10). Copenhagen, Denmark: Association for Computational Linguistics.

Friedrich, A., & Palmer, A. (2014). Situation entity annotation. In *Proceedings of LAW VIII-The 8th Linguistic Annotation Workshop* (pp. 149–158).

Fromknecht, J., & Palmer, A. (2020). UNT Linguistics at SemEval-2020 Task 12: Linear SVC with Pre-trained Word Embeddings as Document Vectors and Targeted Linguistic Features. In *Proceedings of SemEval 2020.* (*to appear*)

Hay, R. J. (2013). Hybrid expressivism and the analogy between pejoratives and moral language. *European Journal of Philosophy*, *21*(3), 450–474.

Hess, L. (2019). Slurs: Semantic and pragmatic theories of meaning. *The Cambridge Handbook of the Philosophy of Language.*

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

Hornsby, J. (2001). Meaning and uselessness: how to think about derogatory words. *Midwest studies in philosophy*, *25*, 128–141.

Kocijan, V., Lukasiewicz, T., Davis, E., Marcus, G., & Morgenstern, L. (2020). A Review of Winograd Schema Challenge Datasets and Approaches. *arXiv preprint arXiv:2004.13831*.

Liu, P., Li, W., & Zou, L. (2019). NULI at SemEval-2019 task 6: Transfer learning for offensive language detection using bidirectional transformers. In *Proceedings of the 13th international workshop on semantic evaluation* (pp. 87–91). Minneapolis, Minnesota, USA: Association for Computational Linguistics.

Mandl, T., Modha, S., Majumder, P., Patel, D., Dave, M., Mandlia, C., & Patel, A. (2019). Overview of the HASOC Track at FIRE 2019: Hate Speech and Offensive Content Identification in Indo-European Languages. In *Proceedings of the 11th Forum for Information Retrieval Evaluation* (p. 14–17). New York, NY, USA: Association for Computing Machinery. Retrieved from `https://doi.org/10.1145/3368567.3368584` doi: 10.1145/3368567.3368584

McCready, E., & Davis, C. (2017). An invocational theory of slurs. *Proceedings of LENLS*, *14*.

Mendelsohn, J., Tsvetkov, Y., & Jurafsky, D. (2020). A Framework for the Computational Linguistic Analysis of Dehumanization. *Frontiers in Artificial Intelligence*, *2*(15).

Newton, C. (2019). *The Trauma Floor: The secret lives of Facebook moderators in America.* https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona. The Verge.

Palmer, A., Robinson, M., & Phillips, K. K. (2017). Illegal is not a noun: Linguistic form for detection of pejorative nominalizations. In *Proceedings of the First Workshop on Abusive Language Online* (pp. 91–100). Vancouver, BC, Canada: Association for Computational Linguistics.

Pandey, A. (2004). Constructing otherness: A linguistic analysis of the politics of representation and exclusion in freshmen writing. *Issues in Applied Linguistics*, *14*(2).

Rahman, J. (2012). The N word: Its history and use in the African American community. *Journal of English Linguistics*, *40*(2), 137–171.

Riggins, S. H. (1997). The rhetoric of othering. *The language and politics of exclusion: Others in discourse*, *8*, 1–30.

Robinson, M. (2018). *A Man Needs a Female like a Fish Needs a Lobotomy: The Role of Adjectival Nominalization in Pejorative Meaning.* Denton, Texas: University of North Texas. Retrieved from `https://digital.library.unt.edu/ark:/67531/metadc1157617/m2/1/high_res_d/ROBINSON-THESIS-2018.pdf` (Master's thesis)

Simon, S., & Bowman, E. (2019). *Propaganda, Hate Speech, Violence: The Working Lives of Facebook's Content Moderators.*

https://www.npr.org/2019/03/02/699663284/the-working-lives-of-facebooks-content-moderators. National Public Radio.

Staszak, J.-F. (2009). Other/otherness. *The International Encyclopedia of Human Geography*.

Toutanova, K., Klein, D., Manning, C. D., & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology* (pp. 173–180).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* (pp. 5998–6008).

Waseem, Z. (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* (pp. 138–142). Austin, Texas: Association for Computational Linguistics.

Waseem, Z., Davidson, T., Warmsley, D., & Weber, I. (2017). Understanding Abuse: A Typology of Abusive Language Detection Subtasks. In *Proceedings of the First Workshop on Abusive Language Online* (pp. 78–84). Vancouver, BC, Canada: Association for Computational Linguistics.

Waseem, Z., & Hovy, D. (2016). Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*. San Diego, California: Association for Computational Linguistics.

Wiegand, M., Ruppenhofer, J., & Kleinbauer, T. (2019). Detection of Abusive Language: the Problem of Biased Datasets. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 602–608). Minneapolis, Minnesota: Association for Computational Linguistics.

Wierzbiecka, A. (1986). What's in a noun? (or: How do nouns differ in meaning from adjectives?). *Studies in Language*, *10*(2), 353–389.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019a). 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)* (pp. 1415–1420). Minneapolis, Minnesota: Association for Computational Linguistics.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019b). SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). In *Proceedings of the 13th international workshop on semantic evaluation* (pp. 75–86).